View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Illinois Digital Environment for Access to Learning and Scholarship Repository

CREATING AN ACCESSIBLE BLOCK ALTERNATIVE FOR THE FOURTH AND EIGHTH GRADE NAEP MATHEMATICS ASSESSMENTS

BY

JEREMIAH M. JOHNSON

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Educational Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2011

Urbana, Illinois

Doctoral Committee:

Professor Lizanne DeStefano, Chair
Professor Hua-hua Chang
Professor Sarah Lubienski
Associate Professor Jinming Zhang

ABSTRACT

For several years NAEP developers and administrators have been interested in creating accessible blocks as a means of improving measurement of student achievement at the lower levels of the NAEP performance continuum, including students with disabilities (SD) and English language learners (ELL). Increased precision at the lower levels represents an important validity issue regarding the use of NAEP as a means of benchmarking and interpreting change in national and state performance over time. The purpose of the current study was to design and field test an accessible block alternative for the grade 4 and grade 8 NAEP math assessments. The study was conducted in two phases. The first phase of the study focused on the development of a set of *Item Modification Guidelines* and *Item Modification Procedures*, and concluded with a small pilot (n = 671 per block). The second phase of the study focused on applying the *Item Modification Guidelines* and *Item Modification Procedures* to create two accessible blocks at each grade level (grade 4 and grade 8), administering the blocks to nationally representative samples of NAEP participants (n = 3,504 for grade 4; n = 3,608 for grade 8), and evaluating the results of the study. Results indicated that accessible blocks significantly reduced estimates of standard error for students at the lower end of the NAEP performance continuum. In addition, results indicated that students who completed an accessible block were significantly less likely to skip items and significantly more likely to complete each item on the assessment. The *Item Modification Guidelines* and *Item Modification Procedures* outlined in this study have been incorporated into the regular NAEP item development and review process.

Two roads diverged in a yellow wood…
And these two roads diverged…
And each of the subsequent roads diverged…
Until, finally, it was realized that
$$\lim_{x \to \infty} 2^x = \infty$$

Each of our roads is unique and unexpected, and, if we are lucky,
we are able to share the road with people who love and care for us.

I would like to dedicate this dissertation to the following people for sharing the road with me:

My father (Glen Johnson) and mother (Renee Schneider) –
For believing in me, caring for me, and showing me how to love.

My advisor (Dr. Lizanne DeStefano) –
For trusting me, teaching me, guiding me, and giving me the opportunity to grow.

My best friend (Joseph Gaines) –
For being there.

My children (Benjamin and Cecilia Quintero Johnson) –
For bringing me joy beyond comprehension.

And

My beautiful wife (Dr. Jessie Quintero Johnson) –
For everything. Forever.

With my whole heart… thank you.

ACKNOWLEDGEMENTS

This study was conducted with the funding and support generously provided by the American Institutes for Research. The National Center for Education Statistics, the Educational Testing Service, Pearson Educational Measurement, and Westat also collaborated to complete this project. All research activities described herein fell under the purview of the NAEP Validity Studies Panel and the National Assessment Governing Board, which is responsible for supervising the development, administration, and reporting of NAEP.

I would also like to acknowledge the many individuals who played a significant role in shaping this work. I owe each a sincere debt of gratitude: Lizanne DeStefano – my advisor and mentor; Fran Stancavage, Phil Esra, Michelle Bullwinkle, Kim Gattis and their colleagues at the American Institutes for Research; Janice Brown and her colleagues at the National Center for Education Statistics; Gloria Dion, Rebecca Moran, Meng Wu, and their colleagues at the Educational Testing Service; Expert item review panel members – Peter Braumfield, Patrick Callahan, Arthur Duvall, Randy McCarthy, Roger Howe, and Wilfried Schmid; Item modification panel members – Theresa Bryant, Jacqueline Bunn, Holly Downs, Aaron Hill, Hsin-Mei Huang, Renee Lemons, Jason Pound, Tony Se, Kathleen Smith, Guy Tal, and Travis Wilson; Cognitive lab administration team members – Maria Jimenez, Shawn Lampkins, and Jessie Quintero Johnson; University of Illinois staff – Linda Morris, Elizabeth Innes, Anne Robertson, and Ronald Banks; And finally, the many students, teachers, principals, and district administrators who agreed to participate in this study.

TABLE OF CONTENTS

CHAPTER 1

NAEP AND ACCESSIBILITY

Assessment of student performance is an important part of the education system and is a crucial catalyst for reform (Stern & Ahlgen, 2002). Since its inception in 1969, the National Assessment of Educational Progress (NAEP) math assessment has served as a leading indicator of student performance and achievement. Today, mounting pressures in education, largely associated with the rise of high-stakes testing and a broad focus on accountability, have created an educational milieu entrenched in local, state, and national student assessment. In this environment, increased attention and importance has been placed on NAEP because it is the only nationally representative, ongoing, and frequent assessment of knowledge of American youth (Berends & Koretz, 1995). NAEP has a reputation for being implemented with a high degree of technical quality, and is considered by many to be the "gold standard" of educational assessment (Daro, Stancavage, Ortega, DeStefano, & Linn, 2007). So being, the agencies and governing bodies responsible for the oversight of NAEP have placed increasing importance on monitoring and improving the validity and reliability of this assessment (Buckendahl, Davis & Plake, 2009b).

For several years, NAEP developers and administrators have been interested in creating accessible blocks as a means of improving measurement of student achievement at the lower levels of the NAEP performance continuum, including many students with disabilities (SD) and English language learners (ELL). To be clear, this interest has been rooted in the belief that NAEP could, and perhaps should, more reliably measure the achievement of these students. Significant numbers of students tend to perform at the

"below basic" level on NAEP. For example, according to the National Center for Education Statistics (NCES), on the 2009 assessment, 18% of grade 4 students performed below the "basic" level in math and only 39% performed at or above the "proficient" level (NCES, 2010). Very small percentages reached the "advanced" level. Furthermore, the percentages of students performing in the lower part of the distribution was much greater for many of the demographic groups that NAEP is required to report by law.

Given the need for NAEP assessments to measure the full range of content and skills specified in the frameworks and achievement level descriptions with relatively few items, the tests have tended to include many items that students find difficult and achievement estimates at the lower extreme of the distribution have had relatively large standard errors (Daro, Stancavage, Ortega, DeStefano, & Linn, 2007). The aim of including one or more accessible blocks would not be to make NAEP easier, but to refine measurement at the lower levels by including more items that provide information about those students' abilities and skills. That is, a primary purpose of including accessible blocks would be to make the NAEP assessment more accessible for low-performing groups of students. The concept of "accessibility" is central to the development NAEP accessible blocks. Appendix A includes a description of accessibility that was included in the 2005 NAEP Mathematics and Item Specifications. Increased precision at the lower levels represents an important validity issue regarding the use of NAEP as a means of benchmarking and interpreting change in state assessment results over time. If state assessment results are showing gains, but NAEP scores remain static for some demographic groups or subject areas, it may be due to NAEP's inability to detect change in the lower performance levels.

The inclusion of an "accessible booklet," consisting of two accessible blocks, also holds promise as a means of increasing the participation of SDs and ELLs and improving the validity of NAEP as a means of representing the performance of those subgroups. Offering an accessible booklet option to SDs and ELLs could be viewed as an accommodation aimed at improving the validity of test results by increasing the amount of assessment information generated for those subgroups and reducing factors that contribute to construct irrelevant variance (e.g., readability, language demand, visual distracters, etc.). This document contains several joint references to SDs and ELLs because, in the context of this study, these populations share two important characteristics. First, the average score for grade 4 and grade 8 SDs and ELLs on the NAEP mathematics assessments is significantly lower than the average score of grade 4 and grade 8 students as a whole. Secondly, SDs and ELLs are frequently excluded from NAEP, while other students are not. It is also acknowledged here that SDs and ELLs are distinct – and quite diverse – student populations with varying educational needs.

Of course, designing accessible blocks suitable for NAEP administration is a formidable challenge. The NAEP is viewed as a "gold standard" in educational assessment, and so being, the agencies and governing bodies responsible for the design, oversight, and conduct of NAEP have established stringent protocols and procedures for item and block development (Kane, 1994). In order to accomplish the task of developing accessible blocks suitable for NAEP administration, two critical (and highly practical) activities must first be completed. First, a definition of what constitutes an accessible block of NAEP math assessment items must be laid out. Second, a process for developing an accessible block of NAEP assessment items that are aligned with the

NAEP content framework(s) and item specifications must be established. Once these activities are completed, it may be possible to develop accessible blocks with a reasonable degree of fidelity.

Purpose

Ultimately, the goal of this study was to advance our understanding of the significance of – and strategies for – improving the accessibility of standardized assessment items. This study describes common threats to the validity of NAEP – and similar standardized assessments – and how improved item accessibility may address some of these concerns. This study also offers guidelines and procedures for increasing item accessibility, and explicates the consequences (i.e., empirical results) of increasing item accessibility in the context of the grade 4 and grad 8 NAEP math assessments. Ideally, this study will be one upon which future scholars, educators, and assessment specialists can expand our understanding of how best to increase the accessibility of standardized assessment items.

More specifically, the purpose of this study was to investigate the feasibility of implementing an accessible block alternative for the grade 4 and grade 8 NAEP mathematics assessments, with a particular emphasis on their potential for increasing precision at the lower levels of the NAEP performance continuum.

An effort to create accessible blocks of NAEP items would be useful for improving the validity of the NAEP assessment in so far as it facilitates or promotes: (a) a reduction of standard error of assessment for populations of students considered the most appropriate candidates for participation in an accessible block (particularly SDs and ELLs), (b) serves to translate the higher level guidance provided by the NAEP framework

into detailed implementation plans for test development, (c) serves to improve the quality (and assurance of quality) for the overall item pool and for individual items, (d) minimizes construct irrelevant sources of item difficulty, (e) expands the range of item difficulty (particularly for students who traditionally underperform on the assessment), and (f) promotes the ideal that NAEP should encompass the achievement of the full population, from the lowest to the highest, and reach from the least to the most advanced content of the framework's domain. These criteria are identical to those laid out by Daro and his colleagues (2007) for the purpose of assessing the validity of the grade 4 and grade 8 NAEP math assessments as a whole.

Research Questions

To begin an exploration of the feasibility of implementing an accessible block alternative, the current study posed the following research questions. These questions were developed to help assess the potential utility of developing an accessible block alternative that is aligned with the current NAEP frameworks and item development processes and standards for the purpose of increasing precision at the lower levels of the NAEP performance continuum.

The first and second research questions relate to understanding the (fundamental) consequences of increasing item accessibility – increased levels of precision, reliability, and student performance – for the student population(s) of interest.

RQ1: How does student performance on modified (accessible) items and unmodified (source) items differ (i.e., to what extent does an accessible block alternative impact item and block percent correct)?

RQ2: To what extent does an accessible block alternative improve the precision/reliability of the NAEP assessment for the lowest performing students (i.e., to what extent are item omission rates and estimates of standard error decreased, and block completion rates increased)?

The third research question relates to understanding the technical quality (and compatibility) of accessible blocks of NAEP items.

RQ3: Can accessible items be scaled along with unmodified NAEP items?

Combined, these questions (and the information provided hereafter) are intended to provide the reader with a reasonable understanding of the relative feasibility and utility of creating and utilizing accessible blocks of NAEP assessment items for the purpose of increasing measurement precision at the lower end of the NAEP performance continuum.

Preview of the Study

The document is divided into five chapters. The second chapter includes a review of literature summarizing current – and persistent – challenges to the validity of the NAEP math assessments. Threats to the validity of NAEP that may be ameliorated by implementing an accessible block alternative are identified. In the third chapter, the design of the accessible block study is explicated. Chapter three also includes a detailed description of the process that was used to create accessible blocks of grade 4 and grade 8 math items. In the fourth chapter, the results of analyses that were completed to address the research questions are detailed. In the fifth chapter, a discussion of the significance of the findings of this study is presented. Chapter five also provides a discussion of the limitations of this study, as well as directions for future research. Six appendices provide supporting documentation.

CHAPTER 2

CHALLENGES TO THE VALIDITY OF THE NAEP

This chapter begins with a brief introduction to the work which has been – and continues to be – conducted to assess and improve the validity of the NAEP math assessments. After this introduction, a brief description of the topic of validity is provided. Challenges to the validity of the grade 4 and grade 8 NAEP math assessments are then discussed. Issues covered in this chapter include: (a) the purpose of NAEP, (b) construct validity (including item quality, construct underrepresentation, and construct irrelevant variance), (c) scoring, (d) standard setting, (e) precision, (f) exclusion, (g) non-participation, (h) unanswered questions (i.e., omitted items), (i) reporting, and (j) accessibility. It is important to understand challenges to the validity of the NAEP math assessments because one can then better evaluate how and why the accessible block alternative can – and cannot – be used to ameliorate these concerns.

Introduction

The NAEP program is subject to constant scrutiny from those who design, implement, and analyze NAEP assessments, and from others with a stake in the educational attainment of U.S. students including policy makers, educational researchers, psychomatricians, members of the media, and various national and international entities (Daro et al., 2007). The validity of the NAEP assessment undergoes regular examination both from within the organizations that oversee and administer the assessment, and from other independent entities. Five sources of operational validity evidence are regularly cited by NCES. These sources include NAEP's Design and Analysis Committee, Task Order Component opportunities, assessment development processes, NAEP-Educational

Statistics Services Institute (NESSI), and the NAEP Secondary Analysis Grant program (Buckendahl, Plake, & Davis, 2009a). Research is also funded through separate programs within the Human Resources Research Organization (HumRRO), the American Institutes for Research (AIR), and the Educational Testing Service (ETS). These validation efforts address a variety of concerns ranging from specific technical or operational questions to broad programmatic questions, and are often built into the normal course of the NAEP development cycle (Buckendahl, Plake, & Davis, 2009a). A significant amount of time, effort, and resources are invested to explore issues related to the validity of the NAEP assessment.

Daro and his colleagues (2007) examined the validity of the grade 4 and grade 8 NAEP mathematics assessments. Some of the validity concerns identified in this report stem from ongoing challenges faced by NAEP that result from its unique position as a national assessment (e.g., Is NAEP unduly oriented toward a particular curriculum, philosophy, or pedagogy?), while others represent broader validity concerns that are relevant to many assessments (e.g., Does the NAEP item pool and assessment accurately reflect the NAEP framework?). This chapter outlines a number of current and ongoing validity concerns for the NAEP grade 4 and grade 8 math assessments, including those that were identified by Daro and his colleagues, and briefly describes the potential utility of creating an accessible block alternative to address some of these concerns.

A Brief Description of Validity

Valid interpretations of scores are a primary concern for any testing program (Buckendahl et al., 2009a). To evaluate the validity of NAEP findings – or any other report of academic performance – a set of key questions must be answered: Can the

results be trusted? Are the results accurate? Are the inferences made from the findings valid and fair? The Standards for Educational and Psychological Testing (American Education Research Association [AERA], American Psychological Association [APA] & National Council on Measurement in Education [NCME], 1999) provides the most authoritative statement of professional consensus of the measurement community regarding the development and evaluation of educational and psychological tests (Linn, 2006). In this document validity is defined as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests," (AERA, APA, NCME, 1999, p. 9). Validation is an ongoing process that should be clear, comprehensive, and explicit. As the stakes for an assessment rise, so too does the requirement for evidence supporting the proposed interpretations and uses of that assessment (Messick, 1990). Similarly, as the importance and relevance of the NAEP assessment has increased, so has the need for continued validation.

An assessment itself is neither valid nor invalid (Buckendahl, Plake, & Davis, 2009a). Evidence of assessment validity should be evaluated in the context of the intended uses and interpretations of the results (AERA, APA & NCME, 1999). Therefore, it is critical that the intended uses and interpretations of NAEP results be specifically identified, and that guidance be provided for gathering evidence to support the validity of the scores for these purposes (Buckendahl, Plake, & Davis, 2009a). Similarly, it is prudent to assist stakeholders in understanding the appropriate and intended uses of NAEP results, as well as inappropriate and unintended uses.

The Purpose of NAEP

Providing a clear statement of purpose is a key step in evaluating the validity of an assessment. The general purpose of NAEP has always been to serve as a broad measure of the status and change in academic achievement of the nation's elementary and secondary students (Tyler, 1966). As the policy and operational agencies respectively, the National Assessment Governing Board (NAGB) and the National Center for Education Statistics (NCES) have consistently indicated that the primary purpose of NAEP is to measure student achievement and change at the national level (Buckendahl, Plake, & Davis, 2009a). However, the specifically stated purpose of NAEP has evolved over time.

The original purpose of NAEP was to monitor student achievement by capturing brief snapshots of the performance of students in U.S. schools across broad regions of the country over time (Tyler, 1966). When Ralph Tyler and other leading technical experts planned NAEP more than 40 years ago, they had hoped that NAEP would inspire the general improvement of the educational system and inform public discourse about education matters. In addition, it was hoped that the publication of sample NAEP items might inspire the professional teaching and testing communities and stimulate the development of new measurement approaches (Baker, 1995). This goal for NAEP was far different from those for tests used to determine the relative proficiency of an individual student (as most state assessments now do). NAEP reports of student performance were intended to reflect on the educational system in general, rather than on individual students or on their schools (Baker, 1995).

In 1983 the Educational Testing Service assumed the responsibility for administering and scoring the NAEP. In 1988, Congress amended the NAEP law to permit state-by-state comparisons and created the National Assessment Governing Board (NAGB), whose task was to decide what students of a certain age *should* know (Bracey, 2009). Subsequently, NAGB released the first version of the NAEP math framework (which greatly assisted in efforts to define and describe the construct being assessed). The NAEP thus became prescriptive as well as descriptive (Baker, 1995).

The purpose of NAEP has since expanded to include the comparison of sub-populations of interest, and even some larger urban districts. The results of NAEP tests are currently used for three major purposes: (a) to monitor trends in student achievement, (b) to provide evaluative statements regarding the level of student achievement, and (c) to make state-by-state comparisons. These purposes are legitimate, that is, supported by legislation and accepted by the general population and education community as reasonable and valuable contributions to our understanding of students' performance in schools. However, these purposes also constitute significant shifts in the stated purpose and accepted interpretations of NAEP data. Evidence of the appropriateness and soundness of these methods and interpretations should be subject to thorough review.

The current uses of NAEP are broadly defined by legislation leaving the actual uses open to a range of interpretation. Although the generic definition offered by NAGB and NCES may intentionally provide flexibility to the agencies that are charged with implementing the program (Buckendahl, Plake, & Davis, 2009a), NAEP's increased visibility in the current educational policy environment raises concerns about the potential for misuse of scores and assessment data. It should also be acknowledged that

there is pressure, whether implicit or explicit, to broaden the purpose of NAEP to include accountability of states for the performance and achievement of students in their schools. Some groups have used NAEP results to benchmark and compare the performance of students in various states and large urban districts, and these comparisons sometimes have significant funding and policy implications for students in those states (e.g., Race to the Top funding applications were frequently supported by references to NAEP). In addition, there have been calls to change the structure of NAEP to provide individual student feedback about their performance against a nationally representative sample of their peers. At this time, NAEP officials only report population and subpopulation estimates of student performance. Current NAEP policies and procedures are intended to ensure that these estimates are as accurate, reliable, and valid as possible.

The purpose of NAEP could be broadened. NAEP is the only national assessment of its kind (i.e., one that endeavors to represent the total population of students in the U.S.). However, the political, societal, and economic ramifications of broadening NAEP's purpose to provide additional information (e.g., student level score reports) should not be underestimated. That is, redefining the purpose of NAEP raises new questions about the validity of the assessment and undermines the validity of the analyses and trends in student performance that NAEP currently reports.

Construct Validity

The NAEP mathematics assessments cover a broad range of knowledge and skills at the grade 4 and grade 8 levels, but do not rest on any specific curriculum or theory of learning (Bracey, 2009). Identifying appropriate learning targets for NAEP assessments is a significant challenge because the target population is very diverse. Many important

issues must be considered when attempting to create coherent, focused assessments in the absence of a shared national curriculum. As a foundational document, the primary purpose of the NAEP mathematics framework is to provide guidance for the development of the NAEP mathematics assessments. The framework is relatively simple and focused, and defines the set of mathematics learning objectives that NAEP endeavors to assess. The current grade 4 and grade 8 NAEP mathematics frameworks are organized into five broad subdomains, or content areas. These content areas are further subdivided (at grade 8) into 20 subtopics and more than 100 objectives, and thus represent a formidable measurement challenge. The item pool used to measure this framework is also ambitious, comprising nearly 170 items at each grade level in 2007 (Daro, et al, 2007).

Construct validity is a central concept underlying assessment validation because the validity of an assessment concerns the meaningfulness of the scores that are reported. The NAEP mathematics framework provides an essential foundation for evaluating the construct validity of the grade 4 and 8 mathematics assessments. The framework provides a standard against which the appropriateness of the content, coverage, range, and balance of items included in the NAEP can be compared. That is, the framework provides both a qualitative and quantitative basis for describing and evaluating the choices NAEP test developers have made. The NAEP assessment is substantially different from most other large scale assessments (especially state assessments) in both purpose and design, and it is therefore difficult to identify benchmarks for how thoroughly the NAEP assessment should, in any given year, cover the content of its framework (Daro et al., 2007).

Three major threats to construct validity include item quality, construct underrepresentation, and construct-irrelevant variance (Pomplun & Omar, 2001). Construct underrepresentation occurs when assessment content is too narrow and fails to include important dimensions of the target construct. Construct-irrelevant variance occurs when assessment content is too broad and contains excess reliable variance associated with other constructs. The presence of irrelevant constructs in the test may result in the test becoming easier or harder for some students in a manner unrelated to the target construct. Item quality, construct underrepresentation, and construct irrelevant variance are discussed in the following portions of this document.

In the context of this study, it is particularly important to understand the relationship between construct validity and item quality, construct underrepresentation, and construct irrelevant variance. Many of the strategies that were employed to create accessible blocks – which are described in this document – were intended to improve item quality and reduce or eliminate sources of construct irrelevant variance. Additionally, "content balance" – a homage to the impact which of construct underrepresentation may have on the validity of an assessment – was used as a criteria for selecting candidate accessible blocks for inclusion in the study.

Item quality. In assessing the validity of NAEP, one must consider the content, alignment, accuracy, and quality of each item in the item pool, of each block of items as a subset of the larger assessment, and of the total item pool for each grade level. Daro and his colleagues (2007) identified two questions that should be addressed when evaluating how well the NAEP item pool aligns to the framework. These questions include: (a) Does each item fit the framework? and (b) How well does the item pool assess the

framework? As a part of their study Daro and his colleagues evaluated the entire grade 4 and grade 8 item pools in which they sought to assess not only the quality of individual items, but also the overall quality of the assessments (i.e., the range, balance, and degree of challenge represented by the item pools as a whole). Daro et al., (2007) defined high quality mathematics items as follows:

> A high-quality mathematics item demands, from the student, knowledge of mathematics and the know-how to reason with mathematics. It does not demand a general ability to decipher complicated presentations or guess what the test maker is looking for. The presentation of the item should be consistent with correct mathematical language available to the student at the grade level being assessed, (p. 77).

This definition stresses the need to ensure that the content of each item in the pool is mathematically accurate, and that the format and presentation of each item are clear, concise, and straightforward. Poorly constructed items may present challenges that include inaccurate or inadequately specified mathematics; unreasonable or hidden assumptions; misleading language, graphics, or contexts; irrelevant complexities; or other cognitive challenges not related to the NAEP framework.

Attention to mathematical quality can produce items that are easy to understand because language is precise and extraneous challenges have been eliminated. However, such efforts can also make items unnecessarily difficult by requiring students to read and comprehend too much explicit information (Rothman, Slattery, Vranek, & Resnick, 2002). Designing and assessing the quality of items written for grade 4 and grade 8 students is not always a straightforward task. Judgments must be made.

Daro et al., (2007) documented examples of "flawed" items in the NAEP item pool, where some aspect of wording, visual display, or context created sources of item difficulty unrelated to the intended mathematical content. Because a significant portion (approximately 30%) of NAEP grade 4 and grade 8 items were found to be “seriously” or “marginally” flawed, Daro et al., suggested that greater mathematics expertise was needed at both the item-writing and review stages of test development. Some items were described as being inconsiderate of the test takers and presented construct-irrelevant challenges that often exceeded the modest mathematical challenge of the item. These irrelevant challenges took the form of poor writing, complicated instruction, misleading presentations, and excessive contexts not related to defining or solving the problem. As troubling as this may seem, the panel concluded that NAEP item quality was virtually the same as a random sample of released state test items, and typical for a large scale assessment. Nevertheless, there is little excuse for having any flaws of this kind on the assessment. NAEP items should exemplify the best in mathematics, not the marginal (Daro et al., 2007).

Construct underrepresentation. Ideally, each NAEP item would be perfectly aligned with the NAEP framework. Alignment is not an attribute of either standards or assessments, but an artifact of the relationship between them. Because alignment describes the match between standards and assessments, it can be improved by altering either one of them, or both (Webb, 1997). Construct underrepresentation is a persistent threat to the validity of the assessment because NAEP is constantly evolving (i.e., the item pool is modified on a regular basis). As a result, the alignment of the NAEP item pool with the targeted objective(s) is also constantly changing. It is incumbent on test

developers to ensure that each item included in the assessment only measure content and skills that are defined in the standards. Similarly, the item pool should fairly and effectively sample the knowledge and skills in the framework, and the assessment should be sufficiently challenging (Rothman, Slattery, Vranek, & Resnick, 2002: Della-Piana, 2008).

Construct irrelevant variance. Every assessment score is made up of the true construct (achievement) plus some amount of construct irrelevant error (Mahoney, 2008). Too much error results in an unreliable assessment, and potentially an inaccurate representation of what a child – or group of children – really knows. This document will briefly address two potential sources of construct irrelevant variance including the use of context in items and language. It is acknowledged that other sources of construct irrelevant variance such as non-uniform testing conditions and student motivation are issues that must be addressed by those that oversee and administer the NAEP assessment. All sources of construct irrelevant variance threaten the reliability and validity of the NAEP math results. However, a complete description of the various sources of construct irrelevant variance is beyond the scope of this study. The use of context and language are highlighted here because these sources of construct irrelevant variance likely affect the scores of students with disabilities (SD) and English language learners (ELL) at a greater rate than the general testing population.

Items presented in context. Many items on the grade 4 and grade 8 NAEP math assessments are presented in some "real-life" context. However, the NAEP mathematics framework does not specify when the use of context is appropriate and/or necessary and when it is not. Clearly, if the intent of a particular item is to assess students' ability to

perform simple computations (e.g., the ability to add fractions), then the use of context may interfere with students' ability to complete the item (i.e., a source of construct irrelevant variance). Likewise, if the intent of an item is to assess students' ability to interpret simple contexts and apply appropriate mathematical concepts, then the use of context would be necessary and appropriate (i.e., a part of the target construct being assessed). There is a difference between embellishing a problem with a context (a practice criticized by mathematicians across the spectrum) and presenting a problem situation out of which the mathematics comes (a practice accepted by mathematicians across the spectrum) (Daro et al., 2007). Without proper guidance, it is often difficult to determine when the use of context is appropriate or necessary, and when it is not.

Language. One must remember that NAEP assessment items are written for children. The demands of mathematical quality must accommodate the demands of communicating with children in the target age range (Rothman, Slattery, Vranek, & Resnick, 2002). The relationship between language proficiency and student achievement on content-based assessments has been well established (Abedi, 2003; Aiken, 1972; August & Hakuta, 1998; Cocking & Mestre, 1988; Kipplinger, Haug, & Abedi 2000; Munro, 1979; NRC, 1999; Orr, 1987; Rothman & Cohen, 1989; Zirkel, 1972). When two constructs function closely together, such as achievement and language proficiency, it is difficult to determine how much of the assessment score is due to true achievement and how much is due to construct-irrelevant variance (Mahoney, 2008).

For tests in English, a student's English proficiency may limit their ability to demonstrate their understanding of the target construct (Robinson, 2010). That is, including English in NAEP math items introduces construct irrelevant variance to the

assessment score as error (to the extent that English is not an intended component of the target construct), which makes interpretations of the assessment scores less reliable and less valid. Standards from the measurement community have cautioned assessment development professionals about the potential validity threats for ELLs taking tests in English (Mahoney, 2008). For non-native English speakers, and for speakers of some dialects of English, every test given in English becomes, in part, a language or literacy test (AERA, APA, & NCME, 1985). Assessment norms based on native speakers of English either should not be used with individuals whose first language is not English; or such individuals' test results should be interpreted as reflecting, in part, current level of English proficiency rather than ability, aptitude, or achievement (AERA, APA, & NCME, 1999). As NAEP policies continue to encourage the increased participation of ELLs, some question the validity of these students' scores because the degree to which their score is a function of their language proficiency is not clearly understood (Rivera & Collum, 2006; Mahoney, 2008). It is important that the scores of ELL students continue to be made publicly available so that their achievement and performance can be compared to that of other student groups.

Scoring

Short and extended response items on NAEP math assessments require human scoring. Before scoring begins, scorers become thoroughly familiar with the items (and relevant standards) they are evaluating. Each scorer reviews and completes each item. For each item, the scorers read and review the associated standards and scoring rubrics, noting aspects of students' responses that are particularly relevant. Scorers then review samples of pre-scored assessment items, and score a preselected sample of student work.

Each reviewer brings their own understanding of mathematics to the task at hand, and attempts to attend to the constraints the students face in completing the items (Rothman, et al., 2002). For example, reviewers are aware of the time allotment and the tools and reference materials that students may or may not use. This familiarization is a critical component of the scoring process.

While NAEP has designed several validity checks into the process of scoring NAEP items, it is clear that a number of judgments must be made about student performance during scoring. It is impossible for item writers and reviewers to foresee the full range of student responses that may be created. Additionally, when scoring rubrics and guides are applied for the first time it is critical for someone familiar with the development of the particular items that are being assessed (including their intended alignment with the NAEP framework and critical components of student responses) be present. This individual should be a valuable member of the team of individuals making decisions and setting precedence for the scoring of individual NAEP items. This is not currently NAEP policy, but could increase the quality and reliability of the scoring process. Scoring guides, once established, become an integral component of the items themselves. The validity of short and extended response items cannot be assessed without also considering the validity of the scoring guides that are used to assess student performance on those items. While common sense and current NAEP scoring standards dictate the manner in which many scoring disputes are to be settled, it is impossible to foresee all possible scoring quandaries that may arise with the development of new items.

Standard Setting

The process of setting cut scores for an assessment is called standard setting. Several different rating methods have been tested in panel studies for the NAEP, but the modified Angoff method has the most solid research base in standard setting, and is currently employed by NAEP administrators (Loomis, 2001). Standards are used to distinguish lower levels of achievement from higher levels of achievement. There are three achievement levels or goals for NAEP: Basic, Proficient, and Advanced. Many students fail to reach a "Basic" level of achievement on the grade 4 and grade 8 NAEP math assessments. These students are commonly regarded as scoring "Below Basic." NAEP's standard-setting process does not begin with a standard then build the test to assess that standard, but rather works the other way around: Standards are established only after the assessments are built (Resnick, 1998). That is, NAEP administrators do not decide in advance what score should be expected of students at different levels of achievement.

Predictably, standards setting is one of the most debated and controversial topics in educational assessment (Cizek, 2001). The process of setting standards to distinguish varying levels of achievement on NAEP has been controversial since the idea was originally proposed (Vinovskis, 1998). The logical arguments and technical evidence undergirding the standard-setting process have been questioned (e.g., Linn, Koretz, Baker, & Burstein, 1991; Stufflebeam, Jaeger, & Scriven, 1991), and defended (Hambleton, Brennan, Brown, Dodd, Forsythe, Mehrens et al., 2000). One concern expressed by various experts was the lack of evidence to support the proposed interpretations of the Basic, Proficient, and Advanced performance levels. A second

concern was that the achievement levels were set too high, causing underestimates of how many students had attained each of the three levels (Pellegrino, 2007). Despite this controversy, it is clear that NAEP achievement level results are one of the most widely regarded indicators of grade 4 and grade 8 students' math achievement (Jaeger, 2003; Zenisky, Hambleton, & Sireci, 2007; Sireci, Hauger, Wells, Shea, & Zenisky, 2009).

Perhaps the fundamental issue in examining the validity of the NAEP standards is whether there is evidence of procedural validity. Evidence of procedural validity is often considered adequate to provide basic support for interpreting performance standards and cut scores unless there is conflicting evidence suggesting that the performance standard or cut score is inappropriate (Kane, 1994). The standard-setting process would not be judged to be valid if there was a lack of evidence of procedural validity, but evidence of procedural validity does not assure the validity of the process (Loomis, 2001). That is, procedural validity is a necessary, but not sufficient, condition for validity.

It is impossible to validate standards or cut scores in an absolute sense (Kane, 1994). The task of evaluating standards involves an assessment of the soundness of the process, and the detection of potential flaws. To support the choice of a performance standard one must show that the cut score is consistent with the proposed performance standard and that this standard of performance represents a reasonable choice, given the overall goals of the assessment program (Sireci, et al, 2009).

Standard setting is a process carried out by reasonable people, and it occurs in a social and political context. As such, it is influenced by multiple factors, including the nature of the assessment itself, current goals and aspirations for the educational

enterprise, practical considerations, sources of comparative data, and immediate social consequences. Collecting validity evidence about the standard-setting process is difficult. Standards are based on judgments. There is no true standard against which to judge the outcome of a standard-setting process (Kane, 1994; Cizek, 2001; Zieky, 2001). However, the NAEP achievement levels serve as a "gold standard" against which other standards are judged, (Kane, 1994). Therefore, NAEP administrators must make every effort to ensure that the statement of the standards is well defined and generally accepted as reasonable. This, in turn, aids in supporting the procedural validity of the standard-setting (i.e., cut score selection) process (Loomis, 2001).

Precision

The precision with which the NAEP estimates the achievement of populations of students depends on a number of characteristics of the assessment. It depends on the number of items administered to each student, and on the degree to which items discriminate among students with different levels of achievement (Allen & Yen, 1979). It also depends on the match of the difficulty of the items to the achievement levels of the students being assessed. Other things being equal, the precision of measurement increases as the number of items administered to each student increases (Daro et al., 2007). Precision is improved when the difficulty of the items are appropriate for the achievement levels of the students being assessed, and when the items have good discriminating power.

It would be relatively easy to design an assessment that would have a high level of precision if the target population for NAEP was narrowly defined. The challenge for NAEP, however, is that the assessment is intended to measure student achievement over

the entire population of assessable grade 4 and grade 8 students. NAEP estimates of state level achievement play an important role in the evaluation of the nation's educational system, and it is important that these estimates have as little error as possible (McLaughlin, Scarloss, Stancavage, Blankenship, 2005). When NAEP administrators report student performance metrics, a standard error is also calculated and reported (McLaughlin et al., 2005). The standard error summarizes the degree of uncertainty in the corresponding statistic. NAEP records the "posterior standard deviation" for each student record, which is an estimate of the size of the error in measuring the student's achievement on NAEP. Although this is not strictly the same as classical standard error of measurement, it is practically equivalent, since it is used in the same way as a standard error of measurement in computing standard errors of aggregate state-level summary statistics.

NAEP estimates of student achievement are based on the performance of a random sample of students, and each student's performance includes inherent random error. These include random errors affecting students' responses to test questions (e.g., carelessness and guessing), the random sampling of students within participating schools, the random sampling of schools within jurisdictions (e.g., sampling error), the random assignment of different test questions to different students under NAEP's "matrix sampling" design, refusal of some sampled schools or students to participate, student absences, imperfections in the lists of all schools from which the NAEP samples are chosen, imperfections in the lists of students at tested grade levels within participating schools, and, for constructed-response items, scorer error (McLaughlin, et al, 2005). The standard error associated with NAEP performance metrics can be reduced either by

increasing the sample size, or by reducing measurement error. Measurement error can be reduced either by increasing testing time, or by assigning students items (or booklets of items) that more closely match their ability level (McLaughlin, et al, 2005).

NAEP test booklets each contain two blocks of items, and these item blocks vary in difficulty. If booklets with at least one accessible block could be administered to the lowest achieving students, then the measurement error for the segment of the population they represent could be reduced. Likewise, if booklets with at least one challenging block could be targeted to the highest achieving students, their measurement error could be reduced (McLaughlin et al., 2005). A savings of 10 percent in the measurement error would produce benefits equivalent to increasing the length of the test or the number of students tested by nearly 20 percent, and a simulation study sponsored by the NAEP Validity Studies Panel estimated that such a reduction should be possible (Linn, Mclaughlin, Jiang, and Gallagher, 2004). Of course, this strategy for reducing measurement error depends on: (a) the availability of NAEP blocks that vary in difficulty from those currently included in the NAEP assessment – such as an accessible block alternative, and (b) the ability of NAEP administrators to develop and implement policies and procedures for identifying and assessing students for whom an alternate assessment (e.g., an accessible booklet) is most appropriate.

The error of measurement varies from student to student, and that variation depends on the "fit" between the student and the test. For the highest achieving students, easy test items provide little information, and for the lowest achieving students, difficult test items provide little information (McLaughlin et al., 2005). The NAEP reporting groups are based on gender, race/ethnicity, eligibility for free or reduced-price lunch,

disability status, and English language learner status. Except for gender, each reporting group constitutes a focal group whose performance distribution is significantly lower than the performance distribution for the population as a whole (Daro et al., 2007). Therefore, measurement precision for these subgroups is differentially affected by the standard error of measurement in the lowest part of the performance continuum. This creates an important validity issue for NAEP, because the performance of these subgroups is often of greatest concern to policymakers. At present, no standard exists on which to judge the significance of the discrepancy in size of standard errors for various populations of interest, but it seems reasonable to be concerned about such a persistent and dramatic pattern that affects those groups of children around which many intervention efforts are focused (Daro et al., 2007).

It is important to note that, over time, NAEP is becoming more and more precise. Additionally, No Child Left Behind mandates for NAEP participation have brought all fifty states into state NAEP for the first time, further increasing sample sizes, and have greatly reduced nonparticipation at the school level (No Child Left Behind Act, 2001; National Assessment of Educational Progress Authorization Act, 2002; Haertel, 2003). The reductions in standard error that can be achieved through the aforementioned techniques (and others) should not be overlooked because decreasing standard error is akin to increasing precision that, in turn, leads to an increase in the reliability and validity of the assessment.

Exclusion

In 2003 a new federal statute required, for the first time, that all sampled schools participate in NAEP. This mandate substantially increased the number of students

participating in the assessment. As NAEP sample sizes have increased, greater precision has been achieved by the program. For this reason, exclusion (i.e., purposefully and systematically removing students from the NAEP sample) effects are increasingly important. Exclusions affect reliability because as students are excluded, the sample size is diminished. As the sample size is reduced the standard error is increased, and the precision of the assessment is reduced (Haertel, 2003). More importantly, NAEP potentially biases the results of the study by excluding a subpopulation of students (Houser, 1995).

A student may be excluded from participating in NAEP either because teachers and test administrators judge that their language fluency is insufficient, or because the student has a disability that would prevent a valid score (Bohrnstedt & Stancavage, 2007). The decision to exclude a student is guided by a process that has been established by NAEP administrators, is supported by NAEP policy, and is implemented by school administrators and teachers (with the support of NAEP representatives) (Bohrnstedt & Stancavage, 2007). Bias can occur in state and national scores if teachers and test administrators do not apply exactly the same exclusion criteria across schools and states. Different exclusion criteria may arise because states have different criteria for classifying students as ELLs or SDs. Additionally, not all states provide the same accommodations for SDs and ELLs. Many states began providing accommodations for their state administered tests before NAEP provided such accommodations. Students accommodated on state tests often were – and continue to be – excluded from NAEP tests because NAEP did not – and in some cases, still does not – offer a similar accommodation (Bohrnstedt & Stancavage, 2007). This is important

because the accessible block could be viewed as an accommodation that, in part, is used to reduce the number students who are excluded from NAEP.

The exclusion rate for a state, district, or demographic group is influenced by two factors. One factor arises from the criteria for exclusion and the way those criteria are actually applied, and the other factor is an artifact of the proportion of students who truly meet those criteria (Haertel, 2003). The estimated bias arising in state NAEP scores from differential and changing exclusion rates across states has been the subject of work completed by the NAEP Validity Studies Panel (McLaughlin, 2000; McLaughlin, 2001). The conclusion of these analysis was that differential and changing exclusion rates could bias scores sufficiently to represent a significant threat to the validity of the scores.

Problems with changing exclusion rates. A portion of students participating in NAEP take an assessment that is designed to provide long-term national time series data. This assessment consists of test items that have not been changed in any of the years that are included in the time series. If the exclusion criteria were altered, NAEP administrators might not be able to make valid comparisons between years for which different criteria were used (Houser, 1995). Since the population being tested would no longer be identical, changes in the time series data could be either the result of actual changes in performance of students or the result of adding more SDs and ELLs to the sample.

In addition to national and state level comparisons over time, differential exclusion rates can affect comparisons among subgroups. Achievement gaps among White, Black, and Hispanic students are an ongoing concern, and attention to gaps and gap reductions has been heightened by the No Child Left Behind Act of 2001. Exclusion

rates vary among subgroups to an even greater extent than among states or districts (Haertel, 2003). Representation of subgroups across states varies considerably as well as the inclusion and exclusion rates for SDs and ELLs. This impacts the validity of the NAEP results for state-by-state comparisons as well as efforts to verify state assessment results (Lane, Zumbo, Abedi, Benson, Dossey, Elliot et al., 2009).

It is impossible to know how well excluded students would have performed if they had been tested, but it is reasonable to assume that, on average, those excluded would have performed worse than those who were actually tested (Haertel, 2003). As "The Nation's Report Card," NAEP is designed to represent the performance of all students at selected grade levels. Exclusions of SDs and ELLs introduces bias in all NAEP statistics, and affects conclusions as to whether changes over time, contrasts among jurisdictions, or differences among subgroups are statistically significant (Haertel, 2003). The effects of exclusions on the reliability of NAEP data can be minimized by: (a) minimizing exclusions, (b) establishing exclusion criteria that are as clear and objective as possible and working to assure that those criteria are adhered to, and (c) making practices and criteria across states as uniform as possible (Haertel, 2003).

Non-Participation

Non-participation can arise either from the absence or refusal to participate of a student chosen in the sample or from a decision of a principal to refuse to allow the school to participate. School participation was voluntary until the 2003 test when participation of sampled schools became mandatory by federal statute (Bohrnstedt, & Stancavage, 2007). However, student participation continues to be voluntary.

Decisions not to participate are made by individuals outside the test administration process, and there is no data collected that would allow exploration of why such decisions are made. Student non-participation is assumed to be caused by a legitimate absence on the day of testing (e.g., illness) unrelated to NAEP testing or by a decision by a parent or student not to participate, which is related to NAEP testing (Bohrnstedt, & Stancavage, 2007). A study by Bohrnstedt & Stancavage (2007) suggests that normal absences are a strong contributor to NAEP student non-participation at both grade 4 and grade 8 and that over one-third of the variance in state student non-participation can be accounted for by normal student absences. This study showed that for grade 4 students in 2003, student non-participation at the national level (5.1%) was only about 1 percentage point above the estimated level of normal absences (4.0%). At grade 8, more than one-half of the non-participation seems attributable to normal absence, and therefore poses little threat to validity. Nevertheless, non-participation may cause (relatively small amounts of) bias in scores, representing a marginal threat to validity, and there is currently no mechanism for collecting data that might be used to explain or describe the population of non-participating students.

Unanswered Questions (Omitted Items)

With the inclusion of short and extended constructed-response items on the NAEP assessments, researchers have begun to notice unacceptably high student non-response (i.e., omit) rates (Koretz et al. 1993, Jakwerth & Stancavage, 2003). Additionally, non-response rates seem to vary with student characteristics like gender and race, and that there are small groups of students for whom omit rates are very high, which may further

impact the validity of NAEP conclusions (Swinton 1991; Zhu and Thompson 1995; Jakwerth & Stancavage, 2003).

Jakwerth & Stancavage (2003) identified several reasons for unanswered questions. These reasons included item format (i.e., constructed response or multiple choice), lack of knowledge/understanding, missed questions (i.e., unintentionally skipped items), motivation, time constraints, test-taking strategy, and testing conditions. Of the item characteristics explored in past studies, only format and difficulty seemed to have any significant relationship with the tendency of an item to be skipped. Studies (Jakwerth & Stancavage, 2003; Koretz et al. 1993; Swinton 1991) have concluded that more open-ended questions tend to be skipped, skipped open-ended questions are often the most difficult, and students seem to stop responding more often at a point where the next question is open-ended rather than multiple-choice.

NAEP policy is to score omitted items as incorrect. However, scoring omitted items as incorrect may result in an underestimation of students' ability (i.e., theta level), particularly for SDs and ELLs. Omitting an item does not necessarily imply that a student does not understand the mathematical concept(s) or principle(s) being assessed in an item. It is also possible that some students choose to omit items because some construct irrelevant feature of the item (e.g., context, language), prevents the student from providing a correct answer. Likewise, it is possible that some items are omitted (especially constructed response items) because students' anticipate that the item will take too much time to complete. Accessible blocks are designed to reduce the likelihood that students will omit items for such reasons.

Reporting

Many constituencies with a stake in the education system use the NAEP as a basis for comparisons (e.g., between states, between subpopulations of interest) and for guidance (e.g., in setting a national agenda for educational reform). There are three primary reasons why NAEP is singled out this way. First, NAEP is the only instrument that endeavors to represent the achievement of all students at the grade 4 and grade 8 levels. Second, NAEP is considered the "gold standard" of educational assessments, largely because of the overall quality of the assessment and the care that has gone into its design and development (Pellegrino, 2007). Third, NAEP performance standards are perceived to have greater rigor and validity than those set for many other assessments, including the achievement tests developed by individual states. Of course, holding the NAEP to such high standards raises several validity concerns for the assessment.

One of the critical tasks in reporting NAEP results is identifying and defining the intended audience for reporting efforts. In 2006 NAGB released a set of *Policy Guidelines* (National Assessment Governing Board, 2006) that explicitly define the primary audience for NAEP as the American public. The guidelines also state that materials used to disseminate NAEP results should be developed for the interested general public, policy makers, teachers, administrators, and parents, and that NAEP results should be distributed to governors and chief state school officers, as well as to superintendents of Trial Urban District Assessment (TUDA) districts. Additionally, national and state organizations with interest in education are notified of NAEP results, and personnel from NCES and NAGB are encouraged to communicate information about

NAEP with various national, state, and local organizations and media representatives (Zeniski, Hambleton & Sireci, 2009).

Over time, NAEP results have been reported with greater levels of detail, to expanded sets of audiences, in order to inform increasingly significant judgments about student achievement and decisions about educational programs. Comparing educational achievement among states and districts, overall and by achievement levels, and disaggregated by major population groups, is far more challenging than national reporting (Noell & Ginsburg, 2009). In addition, NAEP results have been viewed as a basis for comparing state achievement tests results and, appropriately or inappropriately, will probably be used to evaluate standards-setting processes for those states (Pellegrino, 2007). In fact, it appears that NAEP developers themselves may support such uses. In a study conducted by McLaughlin and his colleagues (2005), six states were offered, as an inducement to participate in the study, access to the information on the correlation between their state test and NAEP that would be generated by the study. Although NAEP findings are routinely used to compare states, the validity of those comparisons may be affected by variability among the states in the alignment between state content standards and curriculum and the NAEP assessment frameworks and by inclusion and participation rates for SDs and ELLs (Noell & Ginsburg, 2009).

Even though there have been numerous validity studies to support many of the interpretations and uses of NAEP results, the explication of a comprehensive validity framework to guide the systematic accumulation of validity evidence supporting (or opposing) the proposed interpretations and uses of NAEP results would aid in increasing the validity and reliability the assessment (Lane et al., 2009). Providing clear and

specific statements of the intended and acceptable purposes, uses, and interpretations of NAEP assessment results could also serve to highlight the related issues of fairness and equity. This is of particular importance given the increased use of NAEP to measure the performance of aggregated subgroups (Lane et al., 2009). It is not prudent to leave important decisions about reporting assessment data to the end of the assessment development cycle. The intended purposes, uses, and interpretations of assessment results should be considered throughout the assessment development process, because this would result in the increased likelihood that assessment results are used as intended and are regarded positively (Zeniski, Hambleton & Sireci, 2009).

Important caveats. It is important that NAEP reports include relevant and necessary caveats so that NAEP results can be interpreted in the appropriate light, and so the limitations of the results can be apparent to those that use and interpret those reports (Buckendahl, Plake, & Davis, 2009). Such caveats should include a clear explanation that, although NAEP results are intended to represent the achievement of all students at a particular grade level, the assessment program relies on relatively small samples of students and these students do not take the full assessment. Report writers should also specify that the NAEP assessment frameworks are not specifically aligned (or intended to be) with state frameworks, which are often characterized by content and process standards. Additionally, the numbers that NAEP provides, including both mean scores and percents at-or-above Basic, Proficient, and Advanced are widely interpreted as representing the performance of all students; However, NAEP has never characterized the performance of all students. Exclusions of SDs and ELLs limits the definition of the population being assessed (Haertel, 2003). It should be made clear that exclusions

potentially bias the results of the assessment, and as a result, data are not generalizable to the excluded students (Houser, 1995). In sum, efforts to specify particular interpretations of scores through reporting begins to address the limiting parameters that are often characteristic of defining intended uses, and should be an a standard part of the reporting process (Buckendahl, Plake, & Davis, 2009).

Adjusted v. unadjusted scores. In addition to basic demographic information that is collected from schools, NAEP participants regularly complete a brief background survey. The results of this survey are not used to adjust NAEP scores. Some argue for adjustment, claiming that reporting unadjusted scores without reference to social context differences is inherently unfair, not very informative, and potentially very misleading (Berends & Koretz, 1995; Williams, 1999). Others argue against adjustment, maintaining that adjusting for difference in social context (or reporting group differences along with corollary information about social context) sends an unacceptable message about educational standards. They contend that reporting without adjustment for social-context differences is necessary to communicate that similar expectations are held for all students, not only the privileged.

For certain purposes, reporting only unadjusted differences among population groups may be misleading because these groups tend to come from substantially different family, school, and community contexts, and these contextual differences are in turn powerful predictors of achievement (Berends & Koretz, 1995). White and minority student test score differences that statistically adjust (or control) for dissimilarities in social context are typically far smaller than the unadjusted (raw) population group differences (Berends & Koretz, 1995).

Rank ordering states. NCES regularly rank orders states based on their students' performance on the NAEP assessments. However, Stoneberg (2005) notes that rank order reporting rests on two flawed assumptions about NAEP scores. The first is that each state's score is absolute. NAEP scores, however, are only estimates of state performance determined through a systematic sampling of students and subject matter. As previously noted, not all students in a state are tested, and the students who are assessed do not complete the whole test. The second assumption is that small differences between two NAEP scores justifies ranking one state higher than another. While it is possible to estimate the average achievement of students in a particular state, these estimates are associated with varying levels of measurement error. A rank ordered list of states' performance may identify one state as "outperforming" another when these differences may be statistically insignificant.

Accessibility, Precision, and Participation

The purpose of creating accessible blocks was *not* to make the NAEP assessment easier. The purpose was to make the assessment more *accessible*, particularly for students at the lower end of the performance continuum. Although items included in the accessible blocks were designed and intended to be, in many cases, less difficult than the source item from which they were derived, each of the modified items was scaled with those in the pre-existing NAEP item pool, and was subject to the same item difficulty categorization and classification systems that NAEP currently employs. Of course, increasing accessibility was intended to also increase precision and reliability at the lower end of the performance continuum.

The additional reductions in standard error (i.e., increased precision) that can be achieved by implementing an "accessible booklet option" should not be overlooked. NAEP estimates of state level achievement play an important role in the evaluation of the nation's educational system, and it is critical that these estimates have as little error as possible (McLaughlin et al., 2005). Educational incentive programs such as Race to the Top rely, in part, on NAEP as an indicator of student performance and progress, and as a means for evaluating the relative success of various educational reform efforts. Many states and large urban districts now use NEAP results as tool for establishing benchmarks, identifying student populations of interest or concern, and to inform the discourse relevant to a broad range of issues in education.

Developing and implementing an accessible booklet option could serve to increase the reliability of the NAEP assessment for various demographic subgroups, and to increase the validity and justification for using NAEP results to make important decisions relevant to these groups. As previously noted, NAEP potentially biases the results of the study by excluding some SD and ELL students. It is, therefore, not inconsequential to note that the development and implementation of accessible blocks also holds promise for decreasing the rate of exclusion and increasing student participation. Currently, some SDs and ELLs are excluded from NAEP because standard NAEP assessment booklets cannot adequately represent their abilities and achievement in mathematics. By making available an accessible booklet option, it becomes more likely that NAEP would be able to adequately assess some students' performance, which would result in fewer exclusions. Of course, the successful implementation of an accessible booklet option would require modifications to current NAEP policies and procedures for

accommodating and excluding students. The National Assessment Governing Board and the National Center for Educational Statistics would have to work closely to determine how to utilize accessible booklets to the greatest effect.

Also as previously noted, precision is improved when the difficulty of the items are appropriate for the achievement levels of the students being assessed and when the items have good discriminating power. It depends on the number of items administered to each student, and on the degree to which items discriminate among students with different levels of achievement (Allen & Yen, 1979). The error in students' measurement can be reduced either by increasing testing time, increasing test length, assigning students items (or booklets of items) that more closely match their ability level, or by reducing the construct irrelevant variance that is present in each item. Increasing testing time and length are very expensive, and do little to address validity concerns with the actual assessment items. Accessible blocks are designed to accomplish the more pragmatic goals of including items that more closely match students' ability and reducing construct irrelevant error (to whatever extent possible).

Conclusion

Few (if any) other assessment programs have the scope and substance to influence U.S. educational policy as NAEP can (Zeniski, Hambleton & Sireci, 2009). Although there have been numerous validity studies to support many of the interpretations and uses of NAEP results, the production and implementation of a comprehensive validity framework to guide the systematic accumulation of validity evidence supporting the proposed interpretations and uses of NAEP results should be a priority (Lane, et al., 2009; Noell, & Ginsburg, 2009). This is especially important because NAEP is

increasingly used to monitor the performance of states, districts, and special populations of students.

Undoubtedly, the stakes associated with educational assessments, such as NAEP, are as high as they have ever been. Students, parents, teachers, school and district administrators, and those casting influence on the larger educational system are all attentive to the results of local, state, and national level assessments of student performance. Over the past half century, huge strides have been made in an effort to ensure that the educational needs of all students (including SDs and ELLs) are attended to in schools. Efforts to enhance the overall participation and inclusion of all students in schools are ongoing.

In this context of educational reform, the purpose and value of NAEP is unique. While states are best positioned to develop assessments most appropriate for their own students, there must also be an external benchmark against which to compare the rigor of their standards, tests, and accountability systems. NAEP is that benchmark. Therefore, those responsible for overseeing NAEP to ensure that the assessment is as valid, reliable, and accessible as possible. NAEP is not perfect, and those most familiar with the content and structure of the assessment are aware that it could benefit from greater attention to the lowest performing students. Efforts to develop accessible blocks that are described in this document are intended to serve that purpose.

More than thirty years ago Cronbach (1980) suggested that professional disagreements about interpretations of standards for assessment validity were inevitable. Cronbach noted, "Judgments embody trade-offs, not truths. No matter how much research accumulates, there will be room for divergent interpretations," (p.102). While

this document has described some of the current validity concerns for the grade 4 and grade 8 NAEP math assessments, it should be acknowledged that validity is a broad and complicated concept. In the end, determining the validity of the NAEP assessment is a judgment made by reasonable people who interpret and use the results in a reasonable manner.

Summary

There are many challenges to the validity of the NAEP math assessments. Some of these challenges could be addressed, in part, by implementing an accessible block alternative that was designed to more reliably measure the skills and abilities of the lowest performing students. Validity concerns that may be addressed by incorporating an accessible block alternative into existing NAEP administration policies and practices include: (a) construct validity (including item quality, construct underrepresentation, and construct irrelevant variance), (b) precision, (c) exclusion, (d) unanswered questions (i.e., omitted items), and (e) accessibility. The following portions of this document detail an effort to design, implement, and assess the impact of an accessible block alternative on student performance and the precision with which the achievement of the lowest performing students on could be measured.

CHAPTER 3

METHOD

In this study the research team endeavored to: (a) modify existing blocks of grade 4 and grade 8 NAEP mathematics assessment items in ways that made them more accessible, (b) administer the modified blocks to a random sample of NAEP participants, and (c) compare the performance of students on accessible blocks with the performance of students on source blocks of items.

This study employed multiple methods including expert reviews, cognitive labs, and item modification procedures to create accessible versions of grade 4 and grade 8 NAEP mathematics assessment item blocks that were parallel in purpose and structure to source assessment blocks. After accessible blocks were designed, candidate blocks were selected and administered to a random, nationally representative, sample of NAEP participants (i.e., students who otherwise would have been included in the regular NAEP sample). The results of the study were analyzed to compare the performance of students on the source and accessible blocks. The process of creating, administering, and analyzing accessible blocks of grade 4 and grade 8 NAEP mathematics items was completed in two phases, and that process – including all of the aforementioned research activities – is described in this chapter.

It should be noted that, to the greatest extent possible, this study utilized standard NAEP sampling, administration, scoring, and analysis protocols and procedures. While a complete description of NAEP protocols and procedures is beyond the scope of this document, brief descriptions of relevant NAEP practices are provided as necessary (to provide the reader with sufficient background information to properly interpret the results

of this study). Because the results of this study focus on comparing students' performance on source and accessible versions of the NAEP assessment (with a particular focus on reductions in estimates of standard error for the lowest performing students), a brief overview of current NAEP sampling and psychometric design is provided.

Collaborating groups. The completion of the research activities described in this document required the collaboration of several organizations. The research team worked closely with the American Institutes for Research (who funded the development of the accessible blocks), the National Center for Education Statistics, and the Educational Testing Service during the design, analysis, and reporting phases of this study. The research team also worked with Pearson Educational Measurement during the scoring process. Westat managed all pilot and test administration activities with schools and students, as part of their regular NAEP duties. All research activities described herein fell under the purview of the National Assessment Governing Board, which is responsible for supervising the development, administration, and reporting of NAEP.

Overview of Research Activities

The process of developing and field testing NAEP accessible blocks described in this chapter began in February, 2007 and spanned two full NAEP block development cycles (approximately 4 years). The development of accessible blocks of NAEP math items occurred in two phases. In phase I, two accessible blocks were developed for grade 4 mathematics, and these blocks were evaluated in a 2008 pilot test. The first phase also allowed for the development of the *Item Modification Guidelines* and *Item Modification Procedures*, and provided some initial data about the feasibility and potential utility of designing blocks according to the principles outlined in these documents.

The sample size obtained in the 2008 pilot test was not large enough to scale the items in the accessible blocks. Nevertheless, the results showed that the items were in fact more accessible to students, and thus prompted further efforts to investigate the potential of developing accessible blocks of NAEP math items for grade 4 *and* grade 8 students as a means of reducing measurement error for low-performing students. The second phase of development focused on applying the *Item Modification Guidelines* and *Item Modification Procedures* developed during phase I of the study to create two addition accessible blocks of math items at each grade level using source blocks from the 2009 assessment, administering the blocks, and evaluating the results of the study. The grade 4 item blocks modified in phase II are different from the grade 4 item blocks modified in phase I.

Table 1 below provides a summary of the research activities that were completed during phase I and phase II of the study. Many of the item development activities that occurred during phase I and phase II of the study were similar. This chapter briefly describes each of the research activities that occurred during the study, and important distinctions between phase I and phase II activities are described as necessary and appropriate.

Table 1

Summary of Phase I and Phase II Research Activities

Research Activity	Phase I	Phase II
Experts reviewed source grade 4 NAEP items	X	X
Experts reviewed source grade 8 NAEP items		X
Modified grade 4 source items	X	X
Modified grade 8 source items		X
Drafted *Item Modification Guidelines* and *Item Modification Procedures*	X	
Conducted cognitive labs with two grade 4 candidate accessible blocks (N = 8 per block)	X	
Conducted cognitive labs with four grade 4 and four grade 8 candidate accessible blocks (N = 4 per block)		X
Experts reviewed accessible items, the *Item Modification Guidelines*, and the *Item Modification Procedures*	X	X
Pilot tested grade 4 accessible blocks (N = 671 per block)	X	
Large administration of two blocks per grade level (N = 1,700 per accessible block)		X
Attended scoring session		X
Analyzed data and reported findings to NAEP governing bodies	X	X

It should be noted that there were several important differences between the pilot administration completed during phase I and the larger administration completed during phase II of the study. These differences included: (a) the specific purpose of each administration, (b) the experimental design used to pair accessible blocks with regular

blocks of NAEP items, (c) sampling procedures, and subsequently (d) the population(s) of students assessed.

Previous efforts by ETS to create accessible booklets (consisting of two separate blocks of NAEP assessment items) in math using existing NAEP item blocks had limited success because the performance of the accessible blocks was not sufficiently differentiated from standard blocks (Daro et al., 2007). One reason for this lack of success may have been the absence of a clear and empirically grounded conceptualization of what constitutes an accessible block of NAEP items. Guidelines for item writers that explicitly describe and illustrate how to develop accessible items had also been lacking from previous efforts to develop accessible blocks.

Accessible Block Development

A primary objective of this study was to develop a set of item modification guidelines. The guidelines created for this study were intended to address specific elements of item difficulty (e.g., word choice, item format, graphical constructions, etc.), and their application aided item writers in efforts to develop blocks of accessible items that were aligned with the NAEP frameworks. It should be noted that, throughout the course of this study, the research team treated (and will continue to treat) the item modification guidelines and procedures generated during this study as "living documents" which are subject to continual review and refinement.

The processes of developing an operational definition of a math accessible block (a key objective of phase I of the study) began by convening a panel of content experts with diverse views of mathematics education. This panel was charged with: reviewing the grade 4 and 8 NAEP item pools in math; identifying construct relevant and irrelevant

aspects of the items that contribute to their difficulty; and, offering suggestions for how to make the items easier without compromising the content/construct validity of the items or their alignment with the NAEP framework. Appendix B contains a full list of content expert panel members.

The content expert panel did not directly address item alignment with the NAEP framework. Rather, members of the panel identified factors that increased the difficulty of particular items and proposed strategies for clarifying the measurement intent of the items without altering the construct(s) being measured. The research team then analyzed the item-specific data generated from this process to identify major themes and dimensions that appeared to contribute to item difficulty and to develop general strategies for reducing difficulty without compromising content and construct validity. That is, the expert review process led the research team to develop an initial model for accessible block construction.

Once developed, this working model was reviewed by a second panel composed of four experienced item writers, special education and second language specialists, and math content specialists (i.e., the phase I item modification panel). Appendix C contains a full list of phase I item modification panel members. The phase I item modification panel was asked to further develop the scope, clarity, and potential utility of the item modification guidelines and procedures, and to examine the extent to which the guidelines provided were consistent with the NAEP framework. This task was primarily accomplished in conjunction with the item revision and development process. That is, the phase I item modification panel modified the guidelines and procedures for item modification as they accomplished the task of developing several draft accessible blocks

for review. Appendix D contains the documents that resulted from these efforts, the *Item Modification Guidelines* and the *Item Modification Procedures*.

Expert review of source items. The research team asked a panel of outside reviewers to evaluate the quality of the mathematical content of each of the items in each of the source NAEP blocks included in the study. Each reviewer was a professor of mathematics with expressed interests in mathematics education. Members of the expert panel represent a broad range of mathematical expertise, as well as a broad range of perspectives on mathematics education. Four members of the review panel participated in phase I and phase II of the study. Two additional reviewers were asked to participate in phase II of the study. These new reviewers provided fresh insight into the item review process, and further developed the capacity of the research team to replicate this type of work for future NAEP item review tasks (See Appendix B for a list of expert reviewers).

Each expert reviewer was asked to do three things: (a) rate the mathematical accuracy of every question in each block using the "Item Rating Scale," (b) comment on how well or how poorly each exam was congruent with the NAEP framework, and (c) comment on whether or not each exam was in alignment with the *Item Modification Guidelines*. The Item Rating Scale included three values. A score of 1 meant the quality of the mathematical content presented in the item was adequate. A score of 2 meant the quality of the mathematical content presented in the item was marginal or somewhat problematic. A score of 3 meant the quality of the mathematical content presented in the item was seriously flawed. Please see the Item Rating Scale (Appendix E) for a fuller description of this scale.

During phase I, the initial expert review served as an excellent basis for the item modification process, providing the item modification panel with valuable insights into the mathematical quality and complexity of specific items within the NAEP item pool. During phase II of the study, the initial expert review provided similar, specific, and rich information regarding the mathematical quality of original blocks of items, and once again, this review informed the work of the item modification panel. The review process proved to be a critical step in the process of constructing accessible blocks that were both accessible to the targeted student population(s) and mathematically accurate. It should be noted that the "veteran" reviewers were pleased that many of the general recommendations they had made for improving the NAEP item pool during phase I of the study were reflected in items and blocks under consideration during phase II.

Item modification – phase I. The primary charge of the phase I item modification panel was to examine the feasibility and effectiveness of various strategies for creating accessible blocks that were aligned with the NAEP framework. Because alignment with the NAEP framework and grade level relevance were important, substantial effort was devoted to developing a systematic and rigorous processes for adapting existing NAEP blocks. After factors affecting item difficulty were identified (e.g., item format, "ugly numbers" as opposed to simple (whole) numbers, cognitive complexity), the phase I item modification panel systematically altered seven standard NAEP blocks of grade 4 items, paying close attention to the item modification guidelines and the NAEP framework. The panel then articulated a general process for creating parallel versions of NAEP blocks that offer improved increased levels of accessibility and reductions in construct irrelevant variance (i.e., accessible blocks). A brief description of

this process was amended to the *Item Modifications Guidelines*, and can be found in Appendix D under the heading *Item Modification Procedures*. It was always understood that the application of the *Item Modification Guidelines* via the *Item Modification Procedures* was to be a small part of a larger process for creating accessible blocks. More specifically, it was understood that the process of developing accessible blocks would include: (a) the application of the *Item Modification Guidelines,* (b) initial and final expert reviews of each item in each of the targeted blocks, (c) cognitive labs with purposeful sample of students, (d) field testing of the modified blocks, and (e) extensive review by various NAEP administrators.

The first phase of developing a procedure for item modification involved convening a panel of four item writers and test development specialists. This item review panel created seven accessible blocks of grade 4 math items by adapting "standard" blocks of NAEP assessment items. By systematically varying items in ways intending to increase accessibility and reduce construct irrelevant variance, it was then possible, through the use of cognitive labs and the pilot study, to empirically establish a process for creating accessible blocks of NAEP items. The item modification process occurred over a two-week period in May-June, 2007, and required approximately 70 hours for the item modification panel to complete. During that time the panel completed several tasks including: (a) becoming familiar with the goals of the study, NAEP frameworks, and initial ideas/definition/strategies for creating accessible blocks, (b) examining the feasibility and effectiveness of various processes for creating accessible blocks that are aligned with the NAEP framework while further developing and refining guidelines and recommendations for the creation of accessible blocks, (c) reviewing and adapting seven

existing NAEP grade 4 math blocks by systematically varying items in ways intending to reduce difficulty, (d) developing new items to replace NAEP items that could not be adequately modified, (e) systematically reviewing, editing, and rating each of the seven modified blocks to finalize draft accessible blocks suitable for cognitive lab and pilot testing activities, and (f) providing recommendations regarding which blocks to include in cognitive lab and pilot testing activities. At the end of their work, the item modification panel summarized the procedure they used to modify standard blocks to create accessible blocks, and the resulting document (i.e., the *Item Modification Procedures*) served as the panel's final recommendations for creating future initial drafts of accessible blocks of NAEP math assessment items.

In a limited number of cases, efforts to adapt source NAEP items did not result in an adequate parallel assessment item that was representative of the NAEP framework. In such cases, it was necessary for members of the item modification panel to create new items to round out the accessible blocks and insure alignment with the NAEP frameworks. That is, the current study was guided by a clear and empirically grounded conceptualization of what constitutes an accessible block of NAEP math items. This included identifying elements of difficulty (e.g., item format, language load, graphical constructions, complexity), and systematically varying them in the construction of several pilot accessible blocks. The intention was that the development of math accessible blocks would rely almost exclusively on the adaptation of source NAEP blocks according to the principles outlined in the *Item Modification Guidelines*.

Item modification – phase II. A panel of ten education professionals, math content specialists, individuals with ELL/SD experience, and assessment specialists was

assembled to modify and evaluate each of the items in each of the blocks being considered for inclusion in the 2010 accessible block administration. All potential panel members were interviewed by a member of the research team to assess their strengths and weaknesses (as related to the goals of this study), to determine their availability, and to gauge their level of commitment to the project. All panel members were required to demonstrate a strong understanding of mathematics and/or mathematics education. A total of fifteen individuals were interviewed. Appendix C provides a list of phase II item modification panel members.

During most working sessions, the item modification panel was divided into two teams, with each team concentrating their efforts on a single block of items. Teams of item reviewers were carefully selected so that each team would have a (relative) balance of mathematical, educational, ELL/SD, and assessment expertise. A member of the research team facilitated and closely monitored all aspects of the item modification process. Each block of NAEP items was systematically modified according to the *Item Modification Guidelines* and *Item Modification Procedures* developed during phase I of the study.

The item modification process largely occurred over a four week period during March-April, 2009, and required approximately 80 hours to complete. During this time the item modification panel completed several tasks including: (a) became familiar with the goals of the study, NAEP frameworks, and initial ideas/definitions/strategies for creating accessible blocks, (b) examined the feasibility and effectiveness of various processes for creating "accessible blocks" that are aligned with NAEP frameworks while further developing and refining guidelines and recommendations for the creation of

accessible blocks, (c) reviewed and adapted eleven existing blocks by systematically varying items in ways intended to reduce difficulty and increase clarity, (d) developed new items to replace items that could not be adequately modified, (e) systematically reviewed, edited, and rated each of the eleven modified blocks to finalize draft accessible blocks suitable for cognitive lab activities and NAEP administration, and (f) provided recommendations regarding which blocks to include in cognitive lab and full administration activities.

It should be noted that the item modification panel became more proficient and confident in applying the *Item Modification Guidelines* as their work progressed. Additionally, the item modification panel made minor improvements to the *Item Modification Guidelines* and *Item Modification Procedures* to reflect their thoughts on "best practice" as their work progressed.

The item modification panel carefully recorded and classified each of the modifications that were recommended for each item. Each modification was classified as being either construct relevant (i.e., directly effecting the level or content of the mathematics being assessed) or construct irrelevant (i.e., dealing with issues of format, context, or clarity). Changes to items were considered "construct relevant" if the modification made to the item was likely impact the nature or difficulty of the original task. Figures summarizing the specific modifications made to each of the items at each grade level during phase II of the study are provided below. It should be noted that in a small number of cases the number of alternative answer choices offered for grade 8 items was reduced from five to four. Also, in a small number of cases, the format of an item was changed from constructed response to multiple choice.

Construct Relevant	**75.4%**		**Construct Irrelevant**	**93.0%**
Cognitive Demand	57.5%		Word Choice	51.1%
Graphics	27.7%		Cues	48.9%
Computational Appropriateness	21.3%		Formatting	25.5%
Context	12.8%		Graphics	21.3%
Alternative Answer Choices	10.6%		Alternative Answer Choices	4.3%
Item Format	4.3%		Computational Appropriateness	4.3%
Grade Level Appropriateness	2.1%		Extraneous Information	4.3%
Cues	2.1%		Context	2.1%
Word Choice	0.0%			

Figure 1. Summary of Modifications Made to Grade 4 Items. Percentages based on total number of grade 4 items modified during phase II (n = 47). Three blocks of grade 4 items were modified during phase I and statistics related to these items are not included in this report.

Construct Relevant	**89.6%**		**Construct Irrelevant**	**87.2%**
Cognitive Demand	58.4%		Word Choice	42.4%
Graphics	40.8%		Formatting	35.2%
Alternative Answer Choices	24.8%		Cues	26.4%
Context	24.0%		Graphics	17.6%
Computational Appropriateness	19.2%		Alternative Answer Choices	6.4%
Cues	12.8%		Computational Appropriateness	6.4%
Item Format	2.4%		Extraneous Information	3.2%
Word Choice	0.8%		Context	3.2%
Grade Level Appropriateness	0.0%			

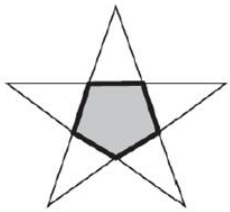
Figure 2. Summary of Modifications Made to Grade 8 Items. Percentages based on total number of items in all modified 8th grade blocks (n = 125).

Some categories appear under both "construct relevant" and "construct irrelevant" headings. These categories include graphics, computational appropriateness, context, alternative answer choices, and word choice. A construct relevant change to a graphic might include adding, deleting, or substantially altering a graphic provided in the original item stem or alternative answer choices. A construct irrelevant change to a graphic might include slight adjustments in graphic placement or content. A construct relevant change

related to computational appropriateness might involve the reduction in the number of mathematical steps required to solve a problem. A construct irrelevant change related to computational appropriateness might involve the elimination of "ugly numbers" from the required calculations. A construct relevant change to context typically involved removing the context of the problem, while a construct irrelevant change to context typically involved simplifying the description of the context. A construct relevant change to alternative answer choices might include (for example) the elimination of an answer choice, or a substantial change to one or more of the alternative answer choices that were originally provided. A construct irrelevant change to alternative answer choices may include (for example) changing the order in which they were presented. A construct relevant change related to cues might involve the provision of a standard formula (e.g., diameter = $2\pi r$), while a construct irrelevant change related to cues might involve bolding or underlining a key word or phrase. A construct relevant change to word choice might involve (for example) changing one or more key words in an item, while a construct irrelevant change to word choice might involve (for example) changing the tense in which an item is presented (past tense to present tense).

For illustrative purposes, two sample grade 4 items that demonstrate the types of modifications that resulted from the application of the item modification guidelines and procedures are included here. The source NAEP items included here were drawn from a pool of ten grade 4 items that were included in the 2007 NAEP administration, and were subsequently released to the public. The accessible items included in this document represent the work of the item modification panel, but were not included in the accessible blocks that were administered to students as a part of this study.

Source Item



What is the shape of the shaded figure inside the star?

A Hexagon
B Pentagon
C Quadrilateral
D Triangle

Accessible Item

What is the shape of the figure?

A Hexagon
B Pentagon
C Rectangle
D Triangle

Figure 3. Sample source and accessible item #1.

Source Item

A loaded trailer truck weighs 26,643 kilograms. When the trailer truck is empty, it weighs 10,547 kilograms. About how much does the load weigh?

A 14,000 kilograms
B 16,000 kilograms
C 18,000 kilograms
D 36,000 kilograms

Accessible Item

A trailer truck weighs 26,843 kilograms. Round this weight to the nearest thousand kilograms.

A 26,000 kilograms
B 26,800 kilograms
C 27,000 kilograms
D 30,000 kilograms

Figure 4. Sample source and accessible item #2.

For the purposes of this study, the modifications made to the star/pentagon figure in and the alternative answer choices in figure 3 would be considered "construct relevant" modifications. In figure 4, the mathematical competencies required to correctly answer the unmodified source item are quite different from those required to complete the modified accessible version of the item. The original item requires subtraction, while the modified version does not. The modified version requires students to correctly identify alternative answer choices that are rounded to the nearest thousand kilograms, while the original item does not. Although the source and accessible items in figure 4 appear to be quite different, they are both designed to assess the same objective which states (in part) that students should , "make estimates appropriate to a given situation with whole numbers, fractions, or decimals," (NAGB, 2004, p.16)

This sample also serves to illustrate the type of feedback received from the expert math panel. The original expert review of this item revealed that many of the math experts were satisfied with the overall validity of the item, but were not at all pleased with the use of the imprecise language "About how much?" Several of the mathematicians noted that this language seemed to indicate that several of the alternative answer choices presented in item one could be perceived as correct (14,000-18,000), as no guidance is provided to the student regarding the degree estimation or rounding that is acceptable.

Selecting candidate blocks. After the process of item modification was complete, members of the item modification panel identified blocks as potential candidates for administration. These blocks were selected based on several criteria including: (a) items within the block were made more accessible while retaining the

integrity of the original testing objective(s), (b) items within the block represented an appropriately diverse range of topics/skills in the NAEP framework, (c) items within the block presented information in multiple ways (e.g., words, pictures, graphs, tables, figures) when appropriate, and (d) the block, as a whole, reflected an appropriate and judicious application of the *Item Modification Guidelines.*

During Phase I, two candidate blocks were selected, and included in cognitive lab and pilot testing activities. During phase II, four candidate blocks at each grade level were selected and included in cognitive lab activities, and two of these four blocks were later selected as top candidates for administration.

Expert review of accessible items. Coinciding with cognitive lab activities, the research team asked the expert review panel to evaluate the quality of the mathematical content of each of the items in each of the accessible blocks of NAEP items. Each reviewer was given the same instructions as were provided during the initial item review. Each reviewer was asked to do three things: (a) rate the mathematical accuracy of every question in each of the eleven exams using the "Item Rating Scale," (b) comment on how well or how poorly each exam was congruent with the NAEP framework, and (c) comment on whether or not each exam was in alignment with the *Item Modification Guidelines*. Again, members of the expert review provided specific, rich information regarding the mathematical quality of modified blocks of items, and their feedback was incorporated into the final versions of the items as appropriate.

Cognitive labs. After general alignment was verified and candidate blocks at each grade level were selected and edited, cognitive labs were conducted to gain insight into how students interpreted and responded to the items and blocks. During the

cognitive labs, both an original block and a parallel accessible block were administered to each student using a counterbalanced design using a 1:1 administration with a trained observer. The observer prompted each student to "think aloud" as they completed the item blocks and debriefed the student as to strategies used once each block was completed. The cognitive lab guide used for this study can be found in Appendix F. Comparisons were made between strategies, time to completion, and performance across accessible and standard blocks.

During phase I, a total of 11 cognitive labs were conducted with grade 4 students. During phase II, a total of 13 cognitive labs were conducted with grade 4 students and 15 cognitive labs were conducted with grade 8 students. All cognitive lab participants were drawn from grade 4 and grade 8 classrooms.

The research team intended to include an overrepresentation of SDs and ELLs in the cognitive labs to investigate the impact of modifications on the performance of these subgroups and the potential use for accessibility purposes. However, a short timeline prohibited the oversampling of these populations.

Selecting blocks for administration. Once cognitive lab and expert review activities were complete, the research team carefully reviewed the available evidence and selected two blocks for administration at each grade level. The research team made every effort to select blocks for administration that represented a judicious application of the *Item Modification Guidelines*, served as a representative sample of the work of the item modification panel, and provided the targeted student population (i.e., students with disabilities and English language learners) with a reasonable chance of demonstrating their skills and abilities relevant to each of objectives targeted in each of the blocks.

NAEP Sample and Psychometric Design

The target population for NAEP includes all students enrolled in public and nonpublic schools in the United States who are enrolled in grades 4, 8, and 12 – and deemed assessable by their school – for the main national NAEP, and ages 9, 13, and 17 for the trend NAEP (NCES, 2003). As previously stated, some of these students are excluded. NAEP is not a simple random sample of students because it does not have a universal sampling frame (i.e., some students are excluded and information provided by schools districts is imperfect); the probability of selection for students differs by subpopulation (i.e., minority students, SD/ELL students and nonpublic school students are oversampled), and sampling units are not independent (i.e., students are clustered in schools) of each other as required by simple random sampling. NAEP has a complex multistage probability sample design (NCES, 2003; Rust & Johnson, 1992). NAEP's national sample is designed so population and subpopulation characteristics (e.g., mathematical ability) can be estimated with a reasonably high degree of precision (Rust & Johnson, 1992). The state NAEP sample is designed with the additional purpose that subpopulation achievement estimates can be obtained with approximately equal precision for all participating states (Rust & Johnson, 1992). Each participating state is required to sample at least 2,500 students from at least 100 schools per subject area (Chromy, 2003).

NAEP samples a large and diverse body of student knowledge (Beaton & Zwick, 1992). To achieve the goal of broad content coverage while minimizing demands on students, NAEP assessments utilize a multiple matrix sampling assessment design (Beaton & Zwick, 1992; Johnson, 1992). Under multiple matrix sampling, each student responds to only a portion of the entire item pool. Beginning with the 1984 assessment,

balanced incomplete block (BIB) spiraling, a variant of multiple matrix sampling, has been used in NAEP assessments. Under BIB spiraling, items at each grade level are first divided into blocks. The blocks are then assembled in to booklets, which contain two blocks of subject items and student background questionnaires. The assignment of blocks of items to booklets is done so that each block appears in the same number of booklets and every pair of blocks appears in at least one booklet. Each block appears in each possible booklet position exactly once (i.e., balanced block positioning). Since no booklet contains all blocks and each student responds to only one booklet, no student responds to all items used to assess a given subdomain (i.e., content area) of the NAEP mathematics framework. In the spiraling stage, the booklets are packaged in a systematic sequence so that each booklet is equally likely to appear in each position in a package. Booklet spiraling ensures that the number of students that receives each booklet is approximately equal. In each testing session, the number of students that receive the same booklet will be small.

Under BIB spiraling design, only a few items in the entire item pool are presented to each student. The number of items within different subdomains of mathematics that are presented to each student is even smaller. This poses problems in the estimation process because it is not possible to reliably estimate an individual student's ability within a given subdomain. Therefore, item response theory point estimates of student ability cannot be used in estimating subpopulation or population distributions (Johnson, 1992). In fact, the use of individually optimal proficiency estimates (such as maximum likelihood estimates) could lead to biased individual estimates in subpopulation or population distributions (Mislevy, Beaton, Kaplan & Sheehan, 1992; NCES, 2003). To

address this problem, NAEP uses plausible value methodology to account for the uncertainty that results from the BIB spiraling matrix sample design.

Due to its complex survey design and oversampling of student subpopulations, analyses of NAEP data incorporate sampling weights so that disproportionate representation of students is accounted for in the estimation process. There are normally four components in deriving student sampling weights (Qian, Kaplan, Johnson, Krenzke & Rust, 2001; Rust & Johnson, 1992). First, a base weight is assigned to a student, which is the reciprocal of the probability of selection. Second, base weights are adjusted for nonparticipation of sampled schools and students. Third, a weight trimming procedure is applied to reduce relatively large weights so that students associated with these large weights will not have an inappropriately large impact on population and subpopulation estimates (Rust & Johnson, 1992). The final stage of student weighting is poststratification. Poststratification ensures that the representation of subpopulations corresponds to the figures from the U.S. census and the Current Population Survey (Braswell et al., 2001; NCES, 2003). Poststratification is also used to reduce mean squared error of estimates associated with student populations that span several subgroups of the population (Qian et al., 2001).

All four components in the derivation of student estimation of weights have their respective functions in the analysis. The base weight and its adjustment for nonparticipation aim at reducing potential bias (Rust & Johnson, 1992). Weight trimming and poststratification reduce sampling error with little introduction of bias (Rust & Johnson, 1992). Because sampling occurs at the school level, students have an unequal probability of being selected for participation in NAEP assessments, and

sampling weights should always be used in computing descriptive statics or conducting inferential procedures. In sum, NAEP has a complex sample design with many special features. Any analysis based on NAEP data should also incorporate these special features.

Item response theory (IRT) models are employed to estimate item parameters (i.e., to scale the items). For dichotomously scored items, the three-parameter logistic (3PL) model is used on multiple choice items. The two-parameter logistic (2PL) model is used on short constructed-response items with two-level rubrics. The generalized partial credit (GPC) model is used for short constructed response items with three-level rubrics and extended constructed response items with four- or five-level rubrics. Using IRT models in scaling has the advantage of invariance of item parameters across different subgroups of students and invariance of ability parameters across different subsets of assessment exercises. IRT scaling is used to produce a common scale where performances of subgroups of students, defined by variables such as gender, race, and disability status, are compared (Braswell et al., 2001).

The sample of students selected for participation in this study was drawn from the pool of 2008 and 2010 operational NAEP participants. That is, a small portion of students in the national sample who would have otherwise completed a regular booklet of assessment items were instead randomly assigned a booklet of items that was assembled for this study. Further information regarding the samples of students draw for the phase I pilot and the phase II administration are provided in the following sections of this document.

Accessible Block Administration

The accessible blocks that were drafted during phase I of the study were pilot tested as part of the 2008 NAEP administration. The research team's ability to incorporate the study as conceived into the 2008 NAEP administration was limited both because the school sample for 2008 had already been drawn and because deadlines for production of materials that were to be included in the 2008 test booklets was fast approaching. ETS, Westat, and Pearson collaborated with NCES to devise a sampling strategy that was feasible, with the intention of enabling NCES, the NAEP Validity Panel, and the research team to realize at least some of the major goals of the study.

As part of the 2008 NAEP administration, accessible booklets of grade 4 math items (consisting of two accessible blocks) were randomly assigned to students using NAEP's standard sampling procedure. For phase I, accessible blocks were paired only with each other, with counterbalancing. This design was intended to maximize data on each accessible block; however, this design also lacked a covariate, as the accessible blocks were not paired with any regular NAEP blocks. The design was intended to yield a per-block sample of at least 600 students. The final sample included 671 students per-block.

This sample was selected to provide critical information regarding the potential value of accessible blocks as a means for more accurately measuring the mathematical abilities of students at the lower end of the NAEP performance continuum (including SDs and ELLs), and of students who are presently excluded from NAEP because of their disability status or level of English proficiency. It was understood that the design and sample of the phase I pilot test would not allow for item scaling or conditioning.

The design of the Phase II administration was significantly more refined. Since there was no regularly scheduled administration of mathematics in 2010, the design for administration and scaling of accessible blocks relied on combining data from the 2010 administration with data from the 2009 operational administration. At each grade level, the 2010 administration included two source blocks: S1 and S2, two accessible blocks: A1 and A2 (where A1 is the modified version of S1 and A2 is the modified version of S2), and two other regular NAEP blocks: S3 and S4. The blocks were arranged in eight booklets, as shown in figure 5 below.

Booklet	Block 1	Block 2
1	A1	A2
2	A2	S1
3	S1	S2
4	S2	A1
5	S4	A1
6	S3	A2
7	A1	S3
8	A2	S4

Figure 5. Block pairings.

Each accessible block thus appeared four times and was paired with every other block except its own source block. Among the regular NAEP blocks, however, the only ones that were paired together were S1 and S2; the rest of the pairings were derived from the 2009 operational data.

A total of 3,000 cases were planned for the 2010 administration at each grade level; a sample size which would provide 375 cases per booklet, 1,500 cases per each accessible item, and 750 cases per each regular NAEP item. The realized sample was slightly larger than required by the design: 3,538 cases at grade 4 (including 372 students with disabilities and 397 English language learners) and 3,608 cases at grade 8 (including 328 students with disabilities and 250 English language learners).

To facilitate item scaling, the sample obtained for phase II of the study was intended to be representative of the larger sample of students who regularly participate in the NAEP assessment. More precisely, students who are normally excluded from participating in the regular NAEP administration were also excluded from the sample selected for phase II of the study. Table 2 and table 3 (below) provide basic demographic information for the sample of grade 4 and grade 8 students that participated in the phase II administration, disaggregated by block and subgroups of interest.

Table 2

Phase II Implementation Sample Demographics by Block – Grade 4

Group	A1 n	A2 n	S1 n	S2 n
Male	877	863	473	487
Female	829	861	452	426
White, not Hispanic	834	840	444	444
Black, not Hispanic	306	329	183	169
Hispanic	427	437	225	225
Asian/Pacific Islander	82	79	45	50
American Indian/Alaska Native	21	15	11	10
Other	36	24	17	15
IEP Yes	142	147	132	126
504 Yes	12	19	10	9
IEP No	1551	1558	783	777
ELL Yes	179	189	105	102
ELL No	1483	1491	798	792
Formerly ELL	43	44	22	18
Total Sample (N)	1706	1724	925	913

Table 3

Phase II Implementation Sample Demographics by Block – Grade 8

Group	n			
	A1	A2	S1	S2
Male	890	925	469	483
Female	897	864	444	422
White, not Hispanic	925	928	477	473
Black, not Hispanic	336	323	161	165
Hispanic	407	409	214	207
Asian/Pacific Islander	83	94	41	41
American Indian/Alaska Native	18	18	9	8
Other	18	17	11	11
IEP Yes	140	143	111	104
504 Yes	17	19	10	12
IEP No	1630	1627	792	789
ELL Yes	106	130	67	68
ELL No	1610	1595	812	806
Formerly ELL	71	64	34	31
Total Sample (N)	1787	1789	913	905

Scoring

As previously noted, short and extended response items included in the accessible blocks required human scoring. While NAEP administrators have designed several validity checks into the process of scoring items, it was clear that a number of judgments were made about student performance during scoring. As a matter of course, it is nearly impossible for item writers and reviewers to foresee the full range of student responses that may be created. It was also clear that, when scoring rubrics and guides were applied for the first time, it was critical for someone familiar with the development of the

particular items that were being assessed (including their intended alignment with the NAEP frameworks and critical components of student responses) be present to be a part of the team of individuals that is making on-the-spot decisions and setting precedence for the scoring of individual NAEP items. This is important because scoring guides, once established, become an integral component of the items themselves. For these reasons, a member of the research team attended the phase II scoring session.

Analysis

After the accessible blocks were developed, administered, and scored, item, block, and grade level analyses were completed. The primary purpose of these analyses was to evaluate the relative success of the item modification efforts. More specifically, efforts were made to: (a) estimate the impact of accessible blocks on student performance (e.g., changes in average percent correct, percent omit, and percent not reached) by block and item for the full population and several sub-populations of interest (e.g., SDs, ELLs), (b) ensure that each item in each of the accessible blocks was scalable with the full NAEP item pool (phase II only), and (c) investigate reductions in standard error of measurement for various levels of student performance (i.e., theta levels) by grade level (phase II only). Each of these data analysis was completed in an effort to address the research questions which guided this study. An overview of each of the major data analysis activities is provided below.

Average percent correct. For each accessible block, the average percent correct was computed and compared to that of the original block of items. If the accessible blocks performed as expected, it was anticipated that the average percent correct for each accessible block, for the full sample as well as each sub-population of interest, would be

significantly higher than the average percent correct for the original block. More specifically, it was hoped that substantial and similar average gains in percent correct (by block) would be observed. Average percent correct by item were also computed and compared to the original item statistics.

Average percent omit. The average percentage of students omitting each item was also assessed. The research team sought to determine if students completing an accessible block tended to omit (i.e., skip) items at a similar rate to students completing a standard block of NAEP items. If items in the accessible blocks performed as expected, it was anticipated that no change or some decrease in the rate at which each item was omitted would be observed. That is, the research team hoped to observe fewer skipped items on the accessible blocks. Average percent omitted by item were also computed and compared to the original item statistics.

Average percent not reached. The average percentage of students failing to reach certain items (i.e., failing to attempt items at the end of each block) was also investigated. If the accessible blocks performed as expected, it was anticipated that students who were given an accessible block would be as likely, or more likely, to reach each item in the block than students who were given the original, unmodified block. Ideally, it was hoped that there would be significant declines in the percentage of students not reaching each item in the accessible block. Average percent not reached by item were also computed and compared to the original item statistics.

Item scaling. Each accessible item was scaled with the full item pool for the NAEP assessment. That is, students' performance on each item, relative to their estimated proficiency (i.e., theta level), was assessed. Item parameter estimates

appropriate for, in most cases, the three-parameter logistic model (including item difficulty, discrimination, and guessing parameters) were computed. The discrimination parameter (i.e., the "a" parameter) represents the degree to which an item discriminates between individuals in different regions on the latent trait (e.g., ability) continuum. The difficulty parameter (i.e., the "b" parameter) serves as an indicator of item difficulty. The guessing parameter (i.e., the "c" parameter) indicates the probability that individuals with extremely low ability will correctly answer the item by chance. If items in the accessible blocks performed as expected, it was anticipated that one would observe little or no change in the average estimate of item discrimination and guessing parameters, by block. More importantly, the research team expected to observe significant reductions in the average estimate of the item difficulty, by block.

Information and standard error estimates. Analyses were completed to determine for which levels of proficiency (i.e., theta levels) each accessible booklet provided the most information. Item response theory promotes the concept of item and test information as an alternative to reliability (Allen & Yen, 1979). Item information is a function of the model parameters, and plots of item information can be used to determine how much information a particular item contributes to the full assessment, and to what portion of the latent trait scale. Because students' scores in individual items are assumed to be locally independent, item information functions are additive. Thus, the test information function is simply the sum of the information functions of the items on the assessment. Test information curves for the original blocks and the accessible blocks were computed and compared. If accessible blocks performed as anticipated, one would

expect that accessible blocks would provided more information for students at the lower end of the performance continuum than the original blocks.

Of course, the amount of information provided by the assessment across the performance continuum is closely related to the estimated standard error of measurement (or more precisely, the conditional standard error of measurement). Because the accessible blocks were designed to provide more information about students at the lower end of the NAEP performance continuum, one would expect to observe an increase in the estimated reliability of students' scores in this range (i.e., a decrease in the observed standard error of measurement for lower-performing students). In fact, the research team had anticipated significant reductions in standard error of measurement on the order of 20-30 percent for these students.

Summary of Methods

In summary, a variety of methods were used to support the development of the accessible blocks during phase I and phase II of the study. Once developed, these blocks were administered to a nationally representative sample of grade 4 and grade 8 participants. Data collected from the phase two administration were analyzed to address the research questions that guided this study.

Alone, none of analyses described above could provide sufficient evidence to assess the relative success of the study. Together, these analyses provided some empirical data about the relative impact of including accessible blocks in the larger NAEP administration. Of course, accessible blocks are intended to increase the reliability of measurement for students at the lower end of the performance continuum. A thorough assessment of the impact of accessible blocks on precision, inclusion, and

validity of assessment results for these students as well as the larger NAEP program should be conducted. Additionally, the implications of the accessible block for NAEP administration, sample size, design, and cost should be considered.

CHAPTER 4

RESULTS

This chapter presents the results of data analysis and is organized by research question. All results presented in this chapter pertain to data collected during the phase II of the study. The primary purpose of this set of analyses was to evaluate the relative success of the item modification efforts. More specifically, efforts were made to estimate the impact of accessible blocks on student performance (i.e., changes in average percent correct), reliability of the assessment (i.e., changes in average percent omit, percent not reached, standard error of measurement, and information), and to determine if each item in each the accessible blocks was scalable. All analyses are presented for both source and accessible blocks by grade, and are disaggregated by SD and ELL status as appropriate.

The Impact of Accessible Blocks on Student Performance (RQ1)

The results presented in this section are intended to address research question 1: How does student performance on accessible items and source items differ (i.e., to what extent does an accessible block alternative impact item and block percent correct)? This section summarizes and compares the performance of students on source and accessible blocks by grade, block, and item. Additional analyses are reported for SDs and ELLs as appropriate.

Percent correct by grade and block. For each accessible block, the average percent correct was computed and compared to that of corresponding source block of items. On average, grade 4 students scored 32.40% higher on the accessible block version of the assessment than the source block version of the assessment, and grade 8 students scored an average of 25.92% higher on the accessible block version of the

assessment than the source block version of the assessment. Table 4 and table 5 below summarize the average change in percent correct by block.

Table 4

Average Percent Correct By Block – Grade 4

Block	N	% Correct
A1	1706	77.23
S1	927	46.27
Difference		+30.96
A2	1726	84.96
S2	914	48.44
Difference		+36.52

Note. Difference was computed by subtracting source from accessible.

Table 5

Average Percent Correct By Block – Grade 8

Block	N	% Correct
A1	1787	75.25
S1	905	49.99
Difference		+25.26
A2	1789	72.41
S2	913	44.19
Difference		+28.22

Note. Difference was computed by subtracting source from accessible.

Similar improvements in student performance were observed for SD and ELL students by block. More specifically, the average shift in percent correct remained relatively consistent regardless of students' disability or English proficiency status.

Table 6 and table 7 below summarize the average percent correct for grade 4 and grade 8 students across the disability categorizations commonly reported by NAEP for each accessible block. Similarly, table 8 and table 9 summarize the average percent correct for grade 4 and grade 8 students across the English proficiency categorizations commonly reported by NAEP for each accessible block.

Table 6

Summary of Percent Correct for Students with Disabilities – Grade 4

Block	IEP	504	No IEP
A1	63.63	73.13	78.48
S1	33.57	37.61	48.48
Difference	+30.06	+35.52	+30.30
A2	74.58	85.06	86.01
S2	38.27	38.50	50.18
Difference	+36.31	+46.56	+35.83

Note. Difference was computed by subtracting source from accessible

Table 7

Summary of Percent Correct for Students with Disabilities – Grade 8

Block	IEP	504	No IEP
A1	53.22	76.42	77.12
S1	32.57	42.81	52.50
Difference	+20.65	+33.61	+24.62
A2	51.49	68.06	74.27
S2	26.78	44.12	46.65
Difference	+24.71	+23.94	+27.62

Note. Difference was computed by subtracting source from accessible

Table 8

Summary of Percent Correct for English Language Learners and Former English Language Learners – Grade 4

Block	ELL	No ELL	Formerly ELL
A1	63.72	78.38	83.03
S1	34.46	47.50	48.38
Difference	+29.26	+30.88	+34.65
A2	76.04	85.82	88.45
S2	38.27	49.43	54.44
Difference	+37.77	+36.39	+34.01

Note. Difference was computed by subtracting source from accessible.

Table 9

Summary of Percent Correct for English Language Learners and Former English Language Learners – Grade 8

Block	ELL	No ELL	Formerly ELL
A1	50.62	76.64	74.85
S1	33.18	51.45	40.24
Difference	+17.44	+25.19	+34.61
A2	52.61	73.95	66.04
S2	24.65	45.83	32.76
Difference	+27.96	+28.12	+33.28

Note. Difference was computed by subtracting source from accessible.

Percent correct by item. For each accessible block and source block, the average percent correct by item was also computed. Figure 6 and figure 7 below summarize the average change in percent correct by item for grade 4 and grade 8

students. Figure 6 summarizes percent correct for all grade 4 items included in the study, and figure 7 summarizes percent correct for all grade 8 items included in the study (i.e., information is combined across blocks at each grade level). Figure 6 shows that, on average, grade 4 participants scored higher the accessible version of items in 29 out of 31 cases. Figure 7 shows that, on average, grade 8 students scored higher on the accessible version of items in 29 out of 32 cases.

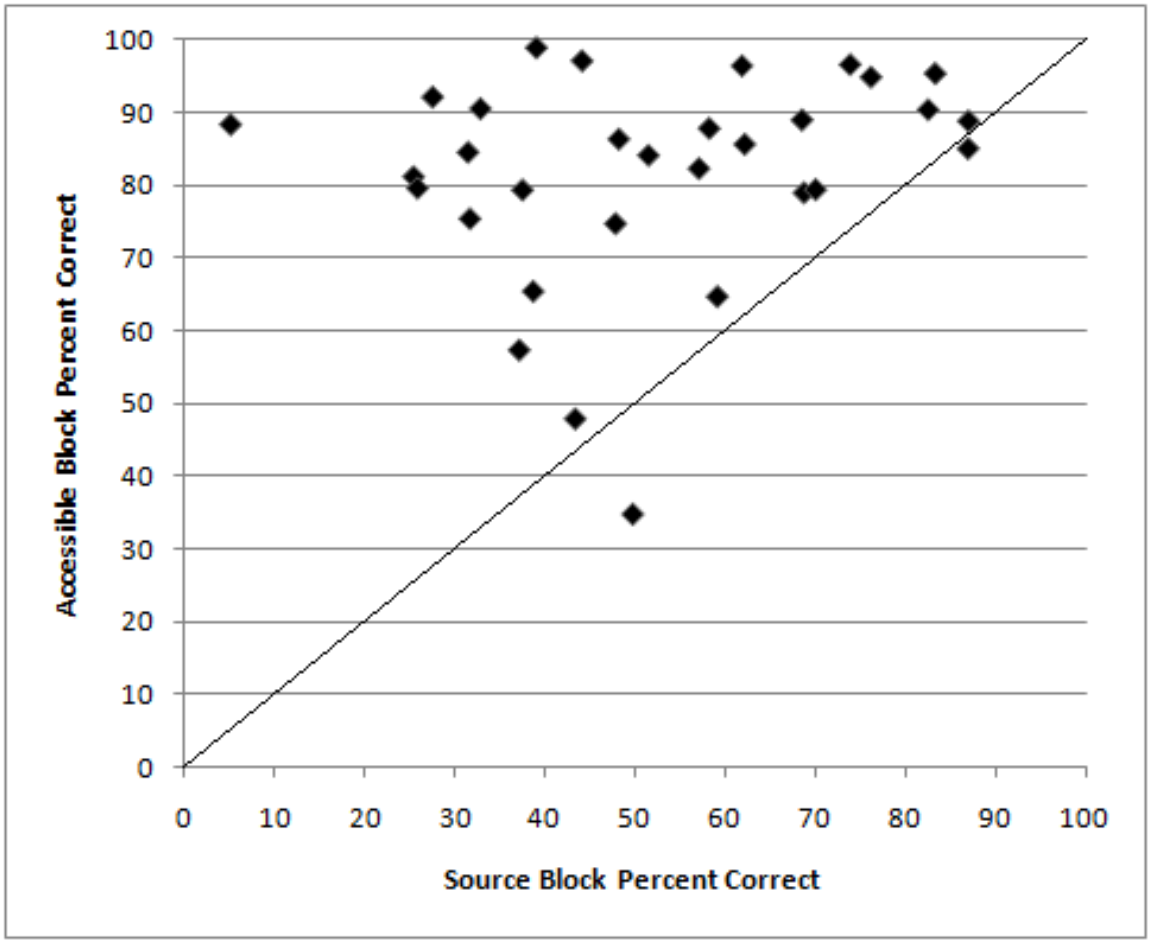


Figure 6. Average percent correct by item – Grade 4.

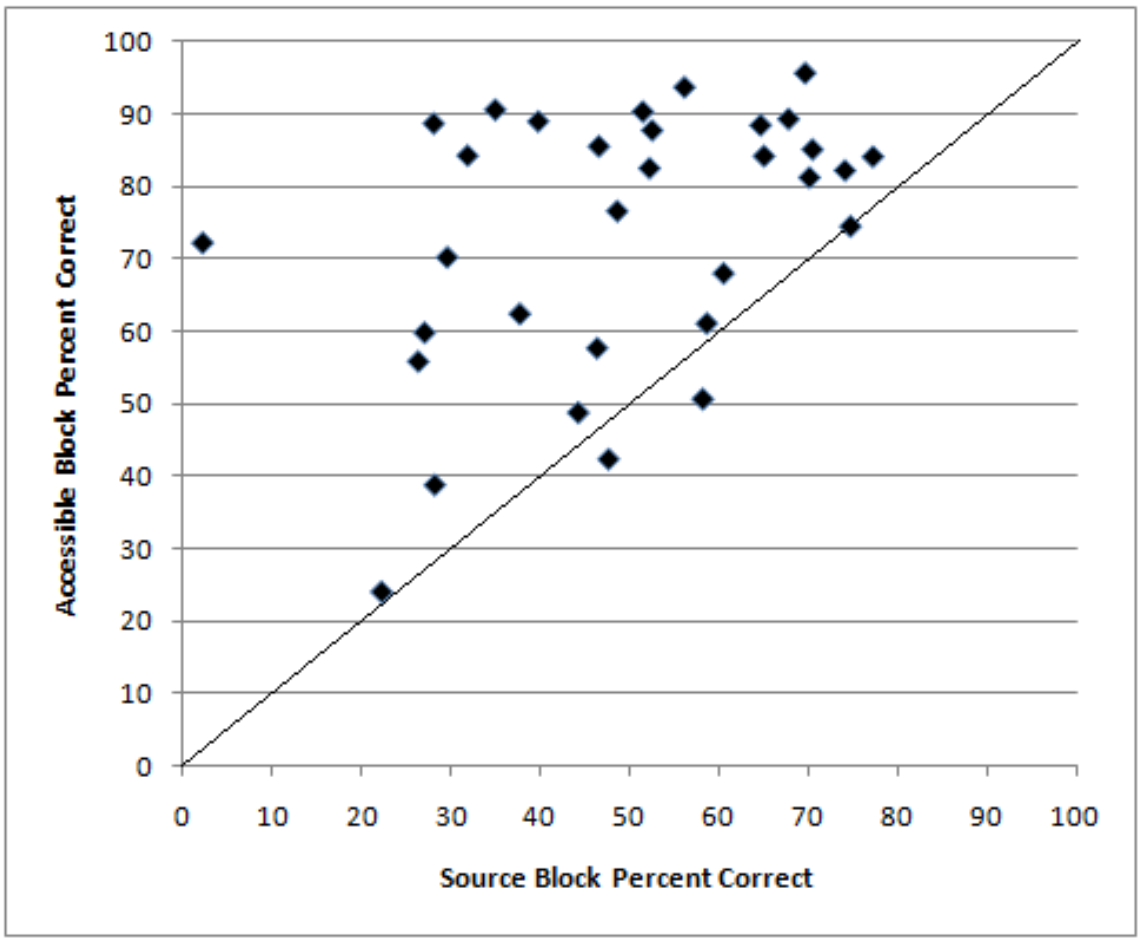


Figure 7. Average percent correct by item – Grade 8.

Item percent correct analyses were also completed for SDs. Figure 8 summarizes percent correct for all grade 4 items included in the study, and figure 9 summarizes percent correct for all grade 8 items included in the study (i.e., information is combined across blocks at each grade level). Figure 8 below shows that, on average, grade 4 SDs scored higher the accessible version of the items in 29 out of 31 cases. Figure 9 shows that, on average, grade 8 SDs scored higher on the accessible version of the items in 27 out of 32 cases.

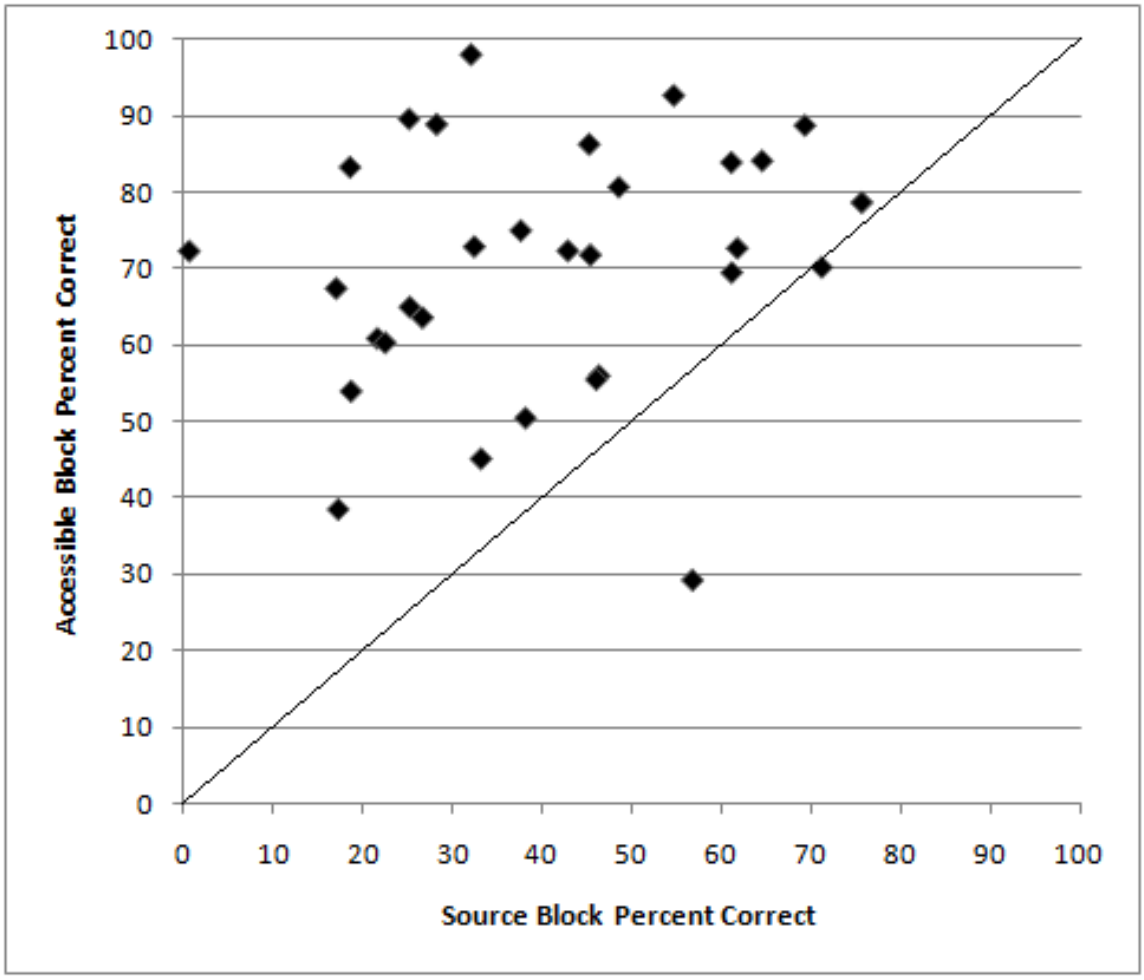


Figure 8. Average percent correct by item – grade 4 students with disabilities.

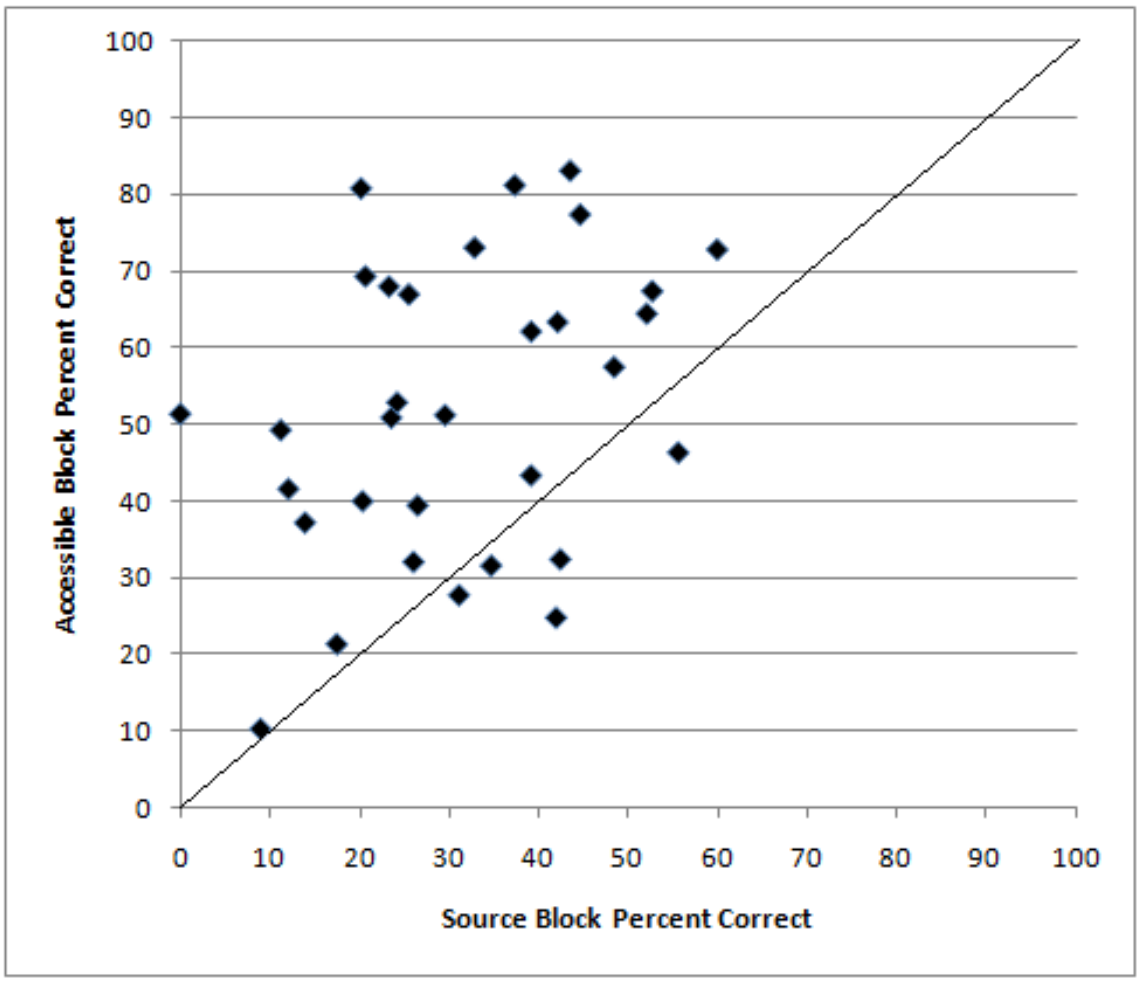


Figure 9. Average percent correct by item – grade 8 students with disabilities.

Similar item percent correct analyses were completed for ELLs. Figure 10 summarizes percent correct for all grade 4 items included in the study, and figure 11 summarizes percent correct for all grade 8 items included in the study (i.e., information is combined across blocks at each grade level). Figure 10 shows that, on average, grade 4 ELLs scored higher on the accessible version of the items in 27 out of 31 cases. Figure 11 shows that, on average, grade 8 ELLs scored higher on the accessible version of the items in 26 out of 32 cases.

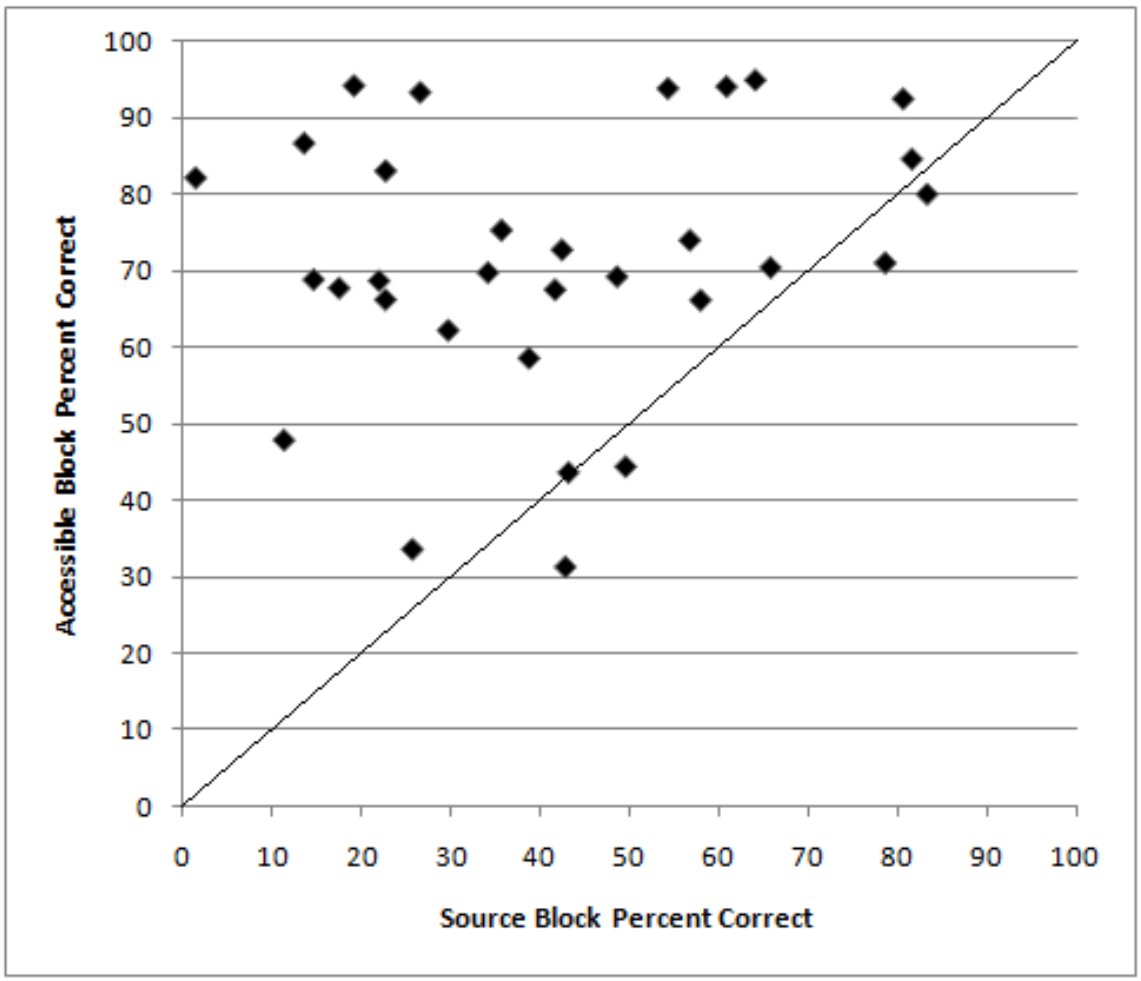


Figure 10. Average percent correct by item – grade 4 English language learners.

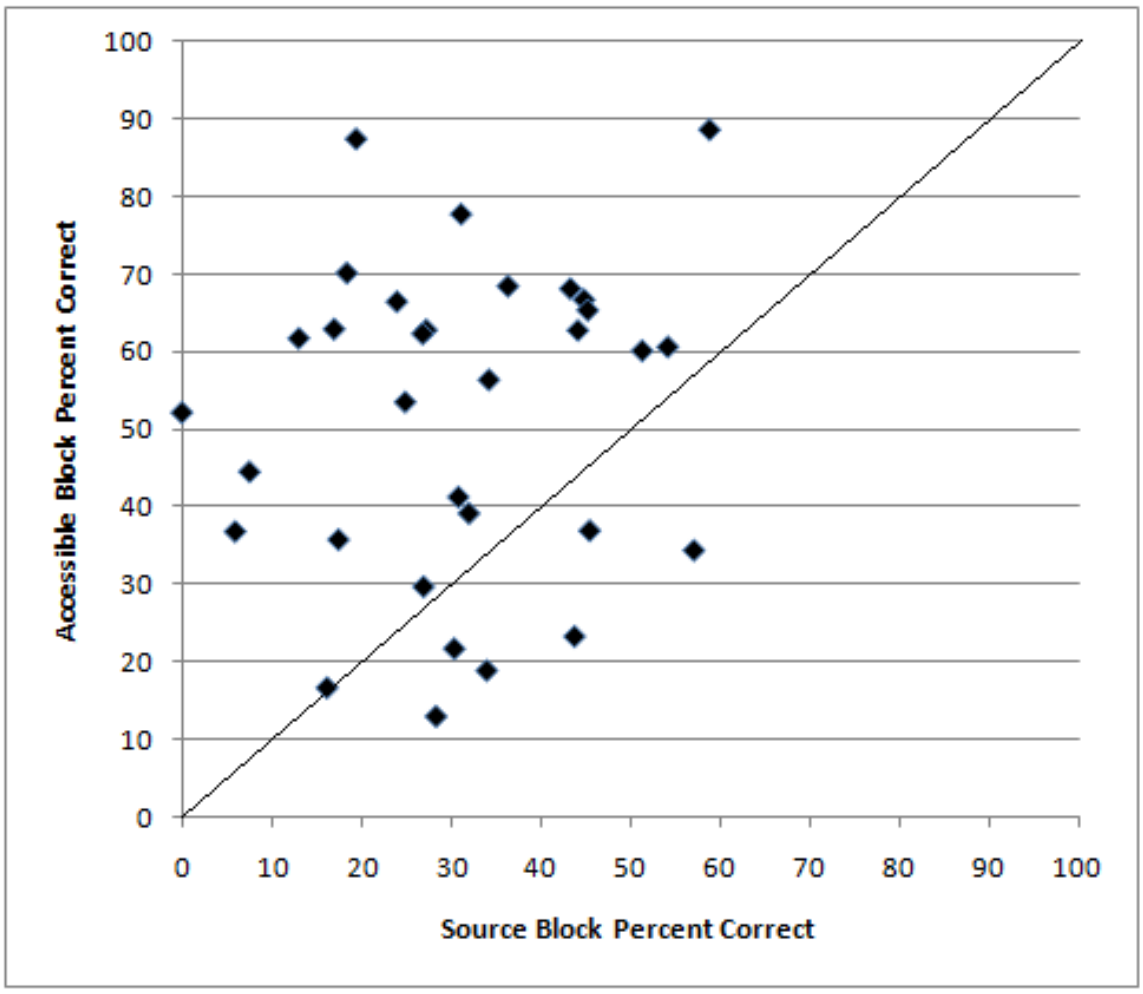


Figure 11. Average percent correct by item – grade 8 English language learners.

The Impact of Accessible Blocks on Precision and Reliability (RQ2)

The results presented in this section are intended to address research question 2: To what extent does an accessible block alternative improve the precision/reliability of the NAEP assessment for the lowest performing students (i.e., to what extent are item omission rates and estimates of standard error decreased, and block completion rates increased)? This section summarizes and compares item omission and block completion rates of students on source and accessible blocks by block, and item. Parallel results for SDs and ELLs are also reported. Estimates of test information and measurement error – across the NAEP performance continuum – are reported by grade.

Items omitted and not reached. For all accessible blocks, a small but significant decrease in the average percentage of omitted (i.e., skipped) items was observed. Additionally, for both grade 4 and grade 8 blocks, there were significant reductions in the average percentage of students not reaching various items on the exam. That is, students assigned an accessible block were more likely to attempt each item in the block than those who were assigned a source block. Table 10 and table 11 below summarize the average change in percent omitted and percent not reached by block.

Table 10

Average Percent Omitted, and Not Reached By Block – Grade 4

Block	N	% Omit	% NR
A1	1706	0.97	0.87
S1	927	1.60	4.24
Difference		-0.63	-3.37
A2	1726	0.59	1.18
S2	914	1.58	5.83
Difference		-0.99	-4.65

Note. Difference was computed by subtracting source from accessible.

Table 11

Average Percent Omitted, and Not Reached By Block – Grade 8

Block	N	% Omit	% NR
A1	1787	0.35	0.61
S1	905	0.98	3.21
Difference		-0.63	-2.60
A2	1789	0.79	1.45
S2	913	1.70	2.55
Difference		-0.91	-1.10

Note. Difference was computed by subtracting source from accessible.

Figure 12 and figure 13 below show the percentage of grade 4 students omitting each item by block type. Combined, these figures show that for grade 4 students the average rate of omission was lower for accessible items than source items in 23 out of 31 cases. Similarly, figure 14 and figure 15 show the percentage of grade 8 students omitting each item by block type. Combined, these figures show that for grade 8 students the average rate of omission was lower for accessible items than source items in 23 out of 32 cases.

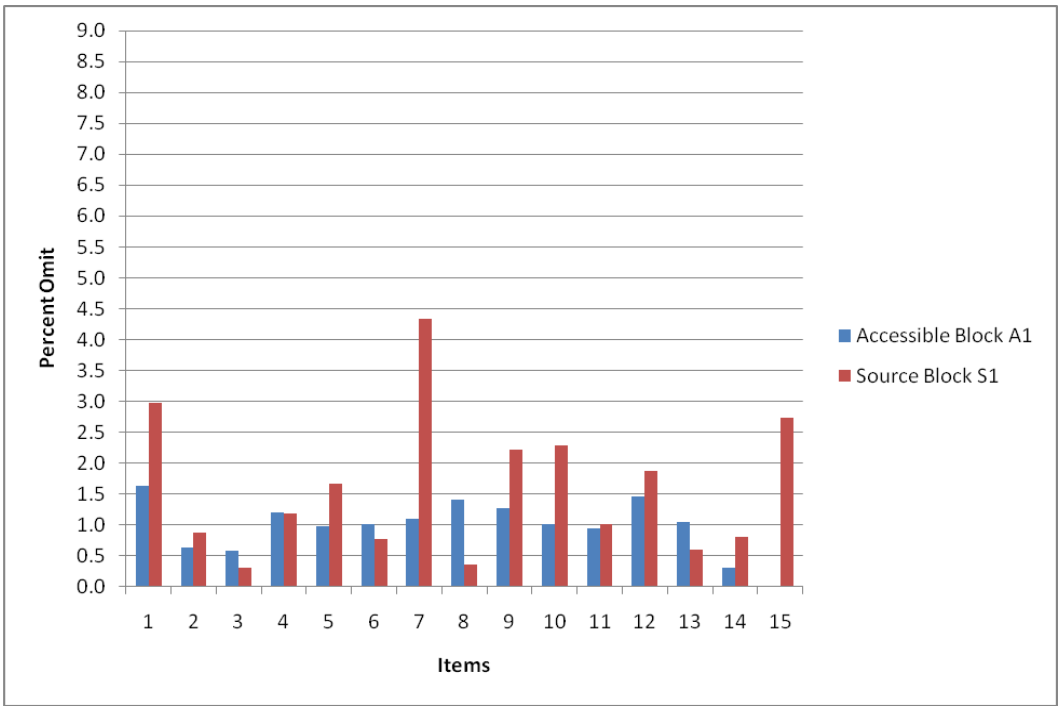


Figure 12. Average percent omitted by item – Grade 4 – Block A1-S1.

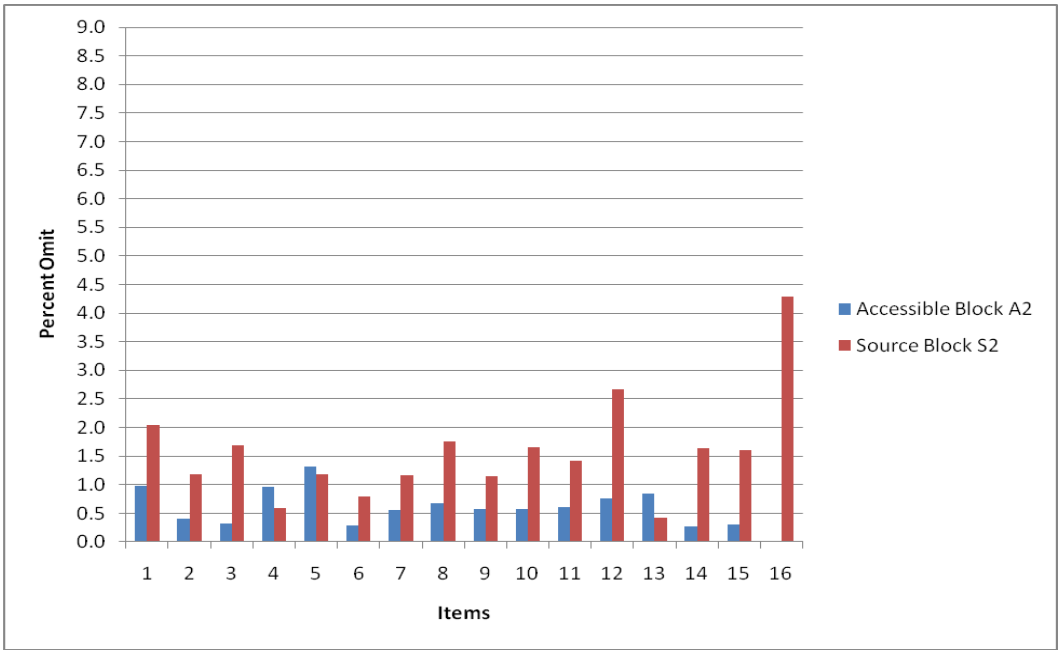


Figure 13. Average percent omitted by item – Grade 4 – Block A2-S2.

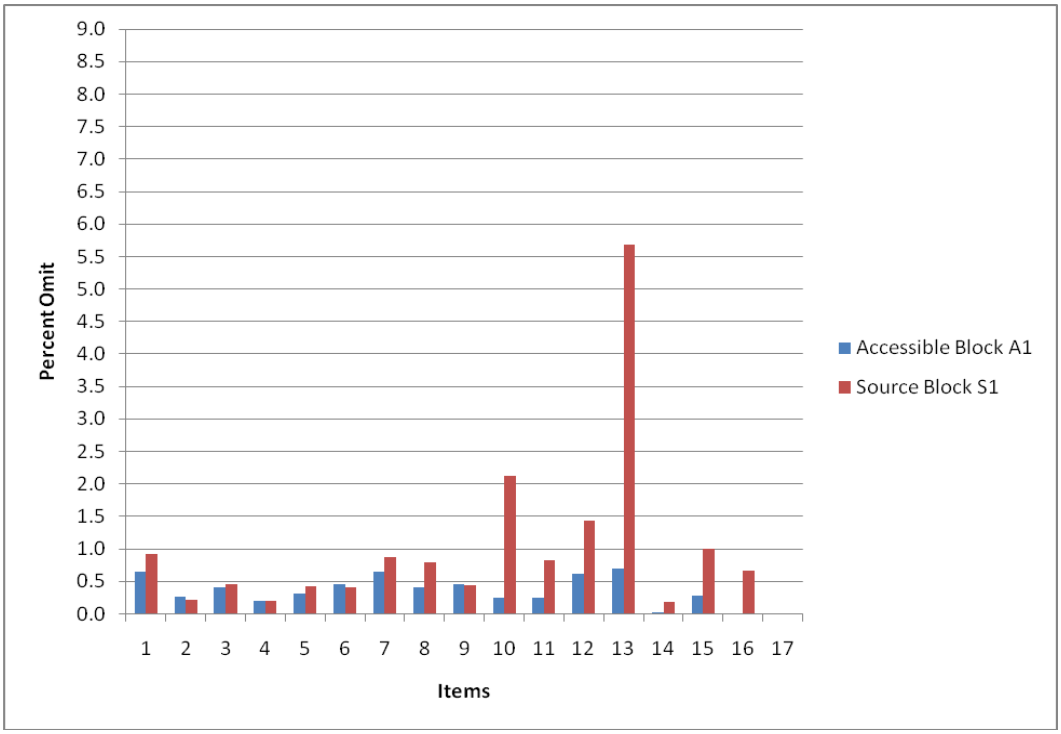


Figure 14. Average percent omitted by item – Grade 8 – Block A1-S1.

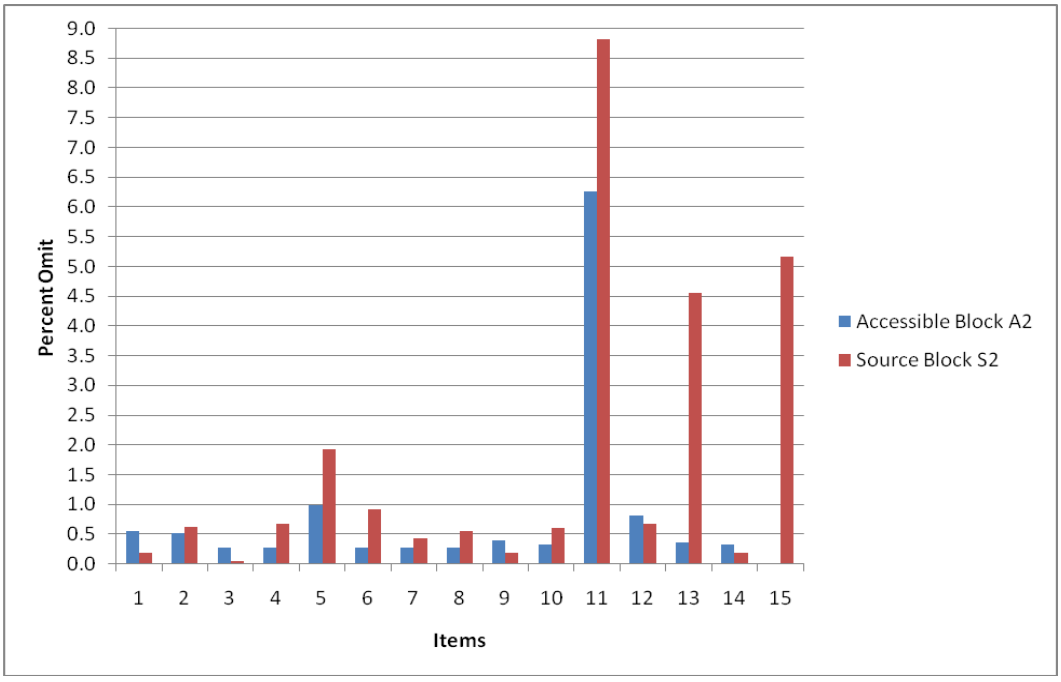


Figure 15. Average percent omitted by item – Grade 8 – Block A2-S2.

Figure 16 and figure 17 below show the percentage show the percentage of grade 4 students failing to reach each item by block type. Similarly, figure 18 and figure 19 show the percentage of grade 8 students failing to reach each item by block type. For all

blocks and grade levels, students failed to reach the last item on the source version of the block more frequently than they failed to reach the last item on the accessible version of the block.

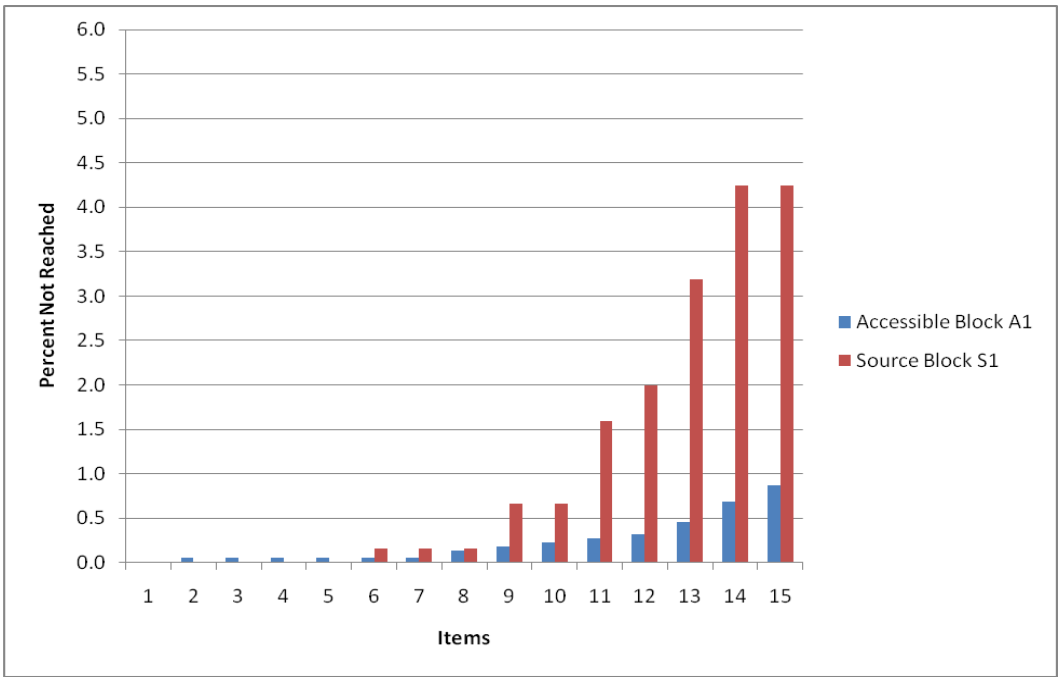


Figure 16. Average percent not reached by item – Grade 4 – Block A1-S1.

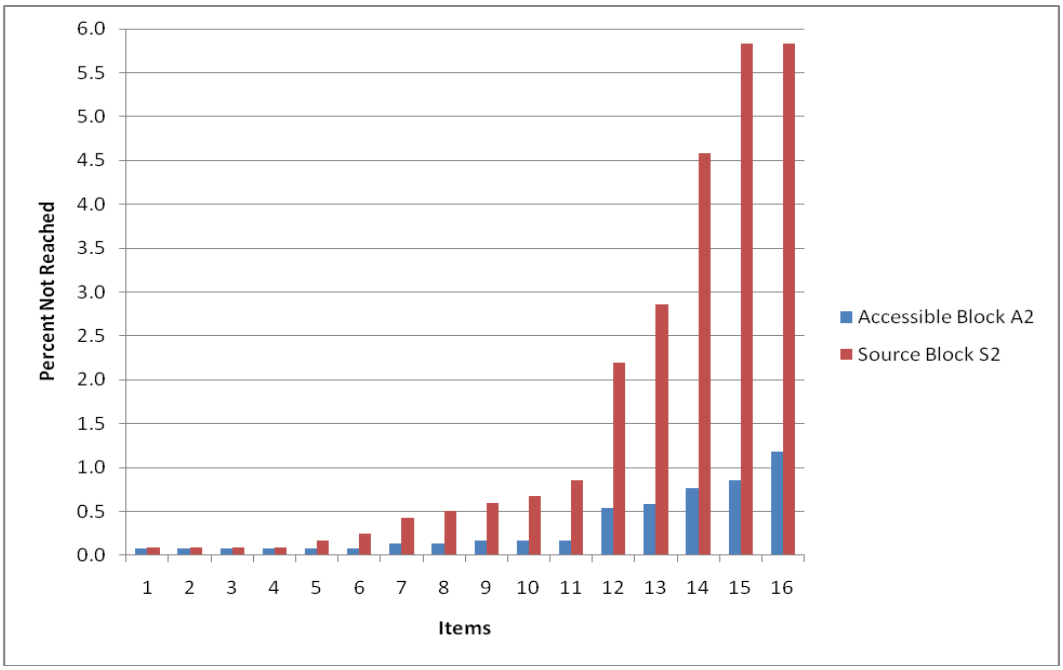


Figure 17. Average percent not reached by item – Grade 4 – Block A2-S2.

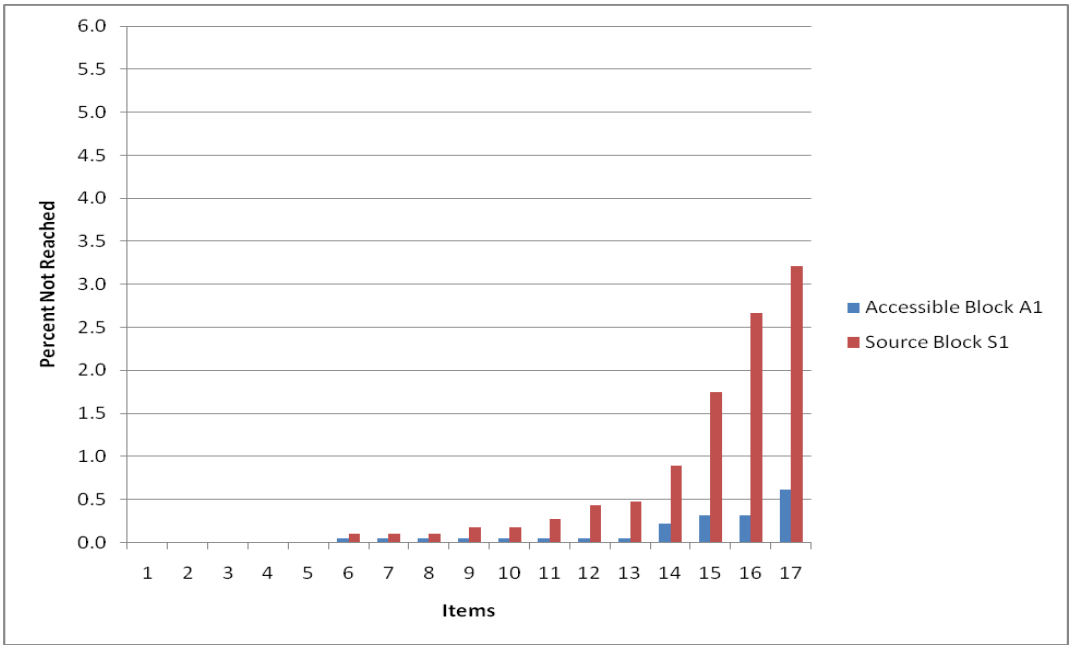


Figure 18. Average percent not reached by item – Grade 8 – Block A1-S1.

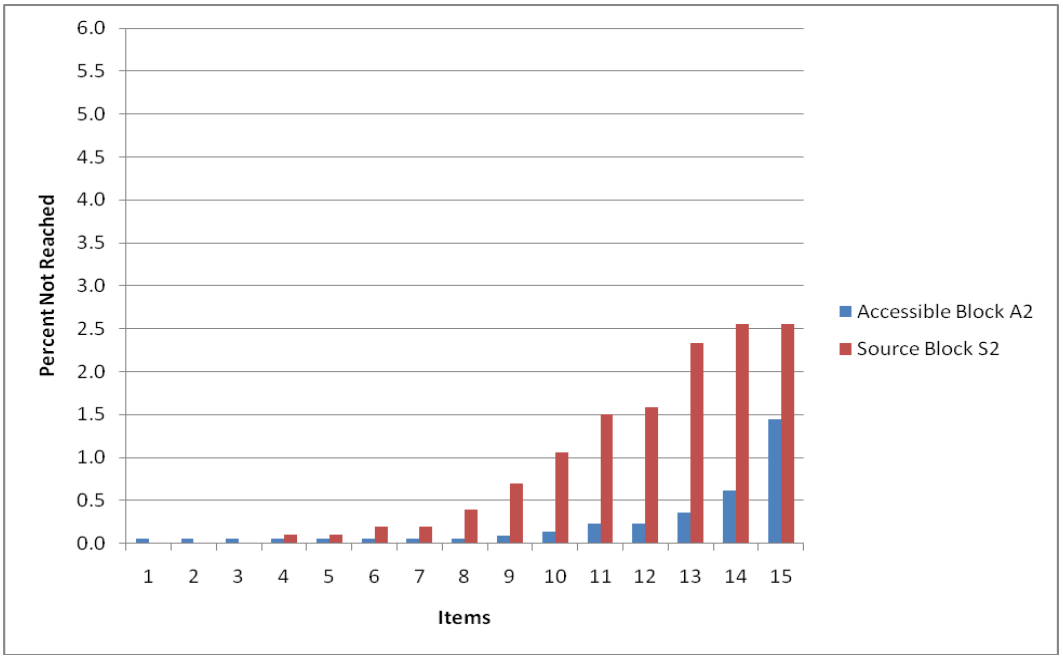


Figure 19. Average percent not reached by item – Grade 8 – Block A2-S2.

Table 12 and table 13 summarize the percentage of grade 4 and grade 8 students omitting various items by disability status. These results are mixed by grade level, block, and disability status. For grade 4 students, those with no IEP were less likely to omit items on the accessible blocks, and students with an IEP were more likely to omit items

on the accessible blocks. For grade 8 students, those with no IEP were less likely to omit items on the accessible block than on the source block. Results for students with a 504 plan are mixed by block and grade.

Table 12

Summary of Percent Omitted by Disability Status – Grade 4

Block	IEP	504	Not IEP
A1	1.61	0.41	0.92
S1	1.39	0.92	1.64
Difference	+0.22	-0.51	-0.72
A2	1.15	0.00	0.54
S2	0.89	0.00	1.69
Difference	+0.26	0.00	-1.15

Note. Difference was computed by subtracting source from accessible.

Table 13

Summary of Percent Omitted by Disability Status – Grade 8

Block	IEP	504	Not IEP
A1	0.55	0.00	0.33
S1	0.55	0.85	1.04
Difference	0.00	-0.85	-0.71
A2	1.06	2.96	0.74
S2	2.84	1.01	1.55
Difference	-1.78	+1.95	-0.81

Note. Difference was computed by subtracting source from accessible.

Table 14 and table 15 below summarize the percentage of grade 4 and grade 8 students omitting various items by ELL status. These results are fairly consistent across

grade, block, and ELL status, and demonstrate that, on average students omitted fewer items on the accessible version of the block than the source version of the block.

Table 14

Percent Omitted by English Language Learner Status – Grade 4

Block	ELL	Not ELL	Formerly ELL
A1	0.60	1.01	0.84
S1	1.26	1.65	0.87
Difference	-0.66	-0.64	-0.03
A2	0.79	0.57	0.36
S2	0.76	1.68	0.55
Difference	+0.03	-1.11	-0.19

Note. Difference was computed by subtracting source from accessible.

Table 15

Percent Omitted by English Language Learner Status – Grade 8

Block	ELL	Not ELL	Formerly ELL
A1	0.76	0.33	0.09
S1	1.86	0.92	1.03
Difference	-1.10	-0.59	-0.94
A2	1.54	0.75	0.69
S2	3.54	1.57	1.97
Difference	-2.00	-0.82	-1.28

Note. Difference was computed by subtracting source from accessible.

Table 16 and table 17 summarize the percentage of grade 4 and grade 8 students failing to reach the last item in a block by disability status. These tables show that across grades, blocks, and disability status, students were more likely to reach the last item in

the accessible block version than the source block version. Item level results for percent not reached by disability status are not reported, but are consistent with previous results.

Table 16

Percent Not Reached by Disability Status – Grade 4

Block	IEP	504	Not IEP
A1	0.58	0.00	0.91
S1	2.97	7.16	4.41
Difference	-2.39	-7.16	-3.50
A2	0.93	3.79	1.17
S2	4.43	0.00	6.12
Difference	-3.50	-3.79	-4.95

Note. Difference was computed by subtracting source from accessible.

Table 17

Percent Not Reached by Disability Status – Grade 8

Block	IEP	504	Not IEP
A1	0.00	0.00	0.67
S1	1.68	5.58	3.38
Difference	-1.68	-5.58	-2.71
A2	1.55	0.00	1.46
S2	3.90	0.00	2.39
Difference	-2.35	0.00	-0.93

Note. Difference was computed by subtracting source from accessible.

Table 18 and table 19 summarize the percentage of grade 4 and grade 8 students failing to reach the last item in a block (i.e., percent not reached) by English proficiency status. These tables show that across grades, blocks, and English proficiency categories,

students were more likely to reach the last item in the accessible version than the source version of the blocks. Item level results for percent not reached by ELL status are not reported here, but are consistent with previous results.

Table 18

Summary of Percent Not Reached by English Language Learners Status – Grade 4

Block	ELL	Not ELL	Formerly ELL
A1	2.24	0.76	0.00
S2	9.23	3.78	0.00
Difference	-6.99	-3.02	0.00
A2	3.96	0.92	0.00
S2	10.35	5.35	5.27
Difference	-6.39	-4.43	-5.27

Note. Difference was computed by subtracting source from accessible.

Table 19

Summary of Percent Not Reached by English Language Learner Status – Grade 8

Block	ELL	Not ELL	Formerly ELL
A1	0.00	0.67	0.00
S1	7.14	3.03	0.00
Difference	-7.14	-2.36	0.00
A2	5.21	1.15	2.83
S2	8.74	2.05	5.14
Difference	-3.53	-0.90	-2.32

Note. Difference was computed by subtracting source from accessible.

Test information and measurement error estimates. Additional analyses were completed to determine for which levels of proficiency (i.e., theta levels) accessible

booklets provided the most information. That is, test information curves for the source blocks and the accessible blocks were computed and compared. Figure 20 and figure 21 below illustrate the estimated ability distribution for grade 4 and grade 8 students (respectively), and three test information curves are superimposed on those distributions. An estimated test information curve is provided for students who would have completed (a) two accessible NAEP blocks, (b) two source blocks of NAEP blocks, and (c) the information that would have been generated about the student population from the original NAEP assessment as a whole. From these figures it is clear that the estimated information generated for students at the lower levels of the NAEP performance continuum is greater for students completing two accessible NAEP blocks than it is for those completing two source NAEP blocks, or the amount of information generated for the student population by the assessment as a whole.

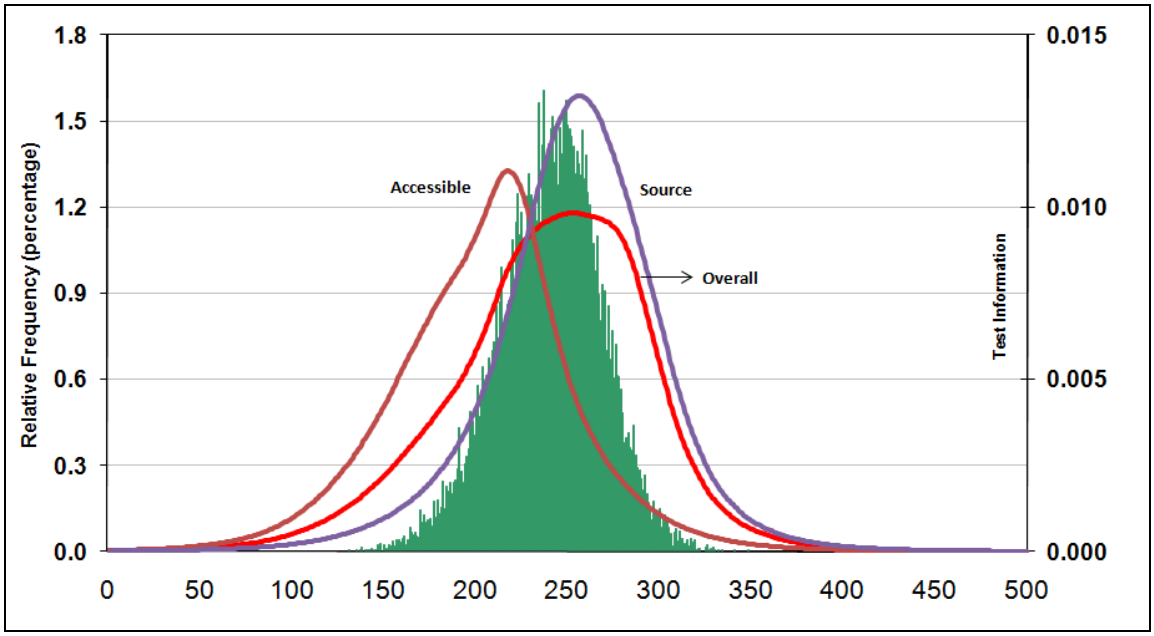


Figure 20. Ability distribution and test information by book type – Grade 4.

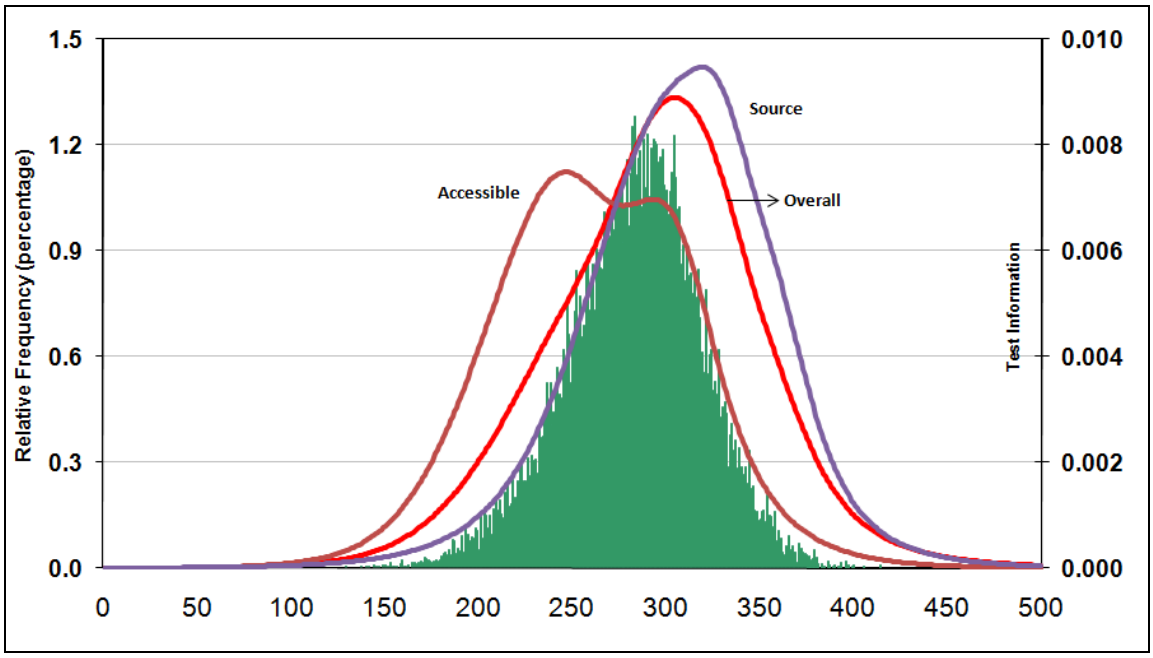


Figure 21. Ability distribution and test information by book type – Grade 8.

Figure 22 and figure 23 below illustrate the estimated ability distribution for grade 4 and grade 8 students (respectively), and three estimated CSEM curves are superimposed on those distributions. An estimated CSEM curve is provided for students who would have completed two accessible blocks, two source blocks, and the estimated CSEM that observed across the student population from the original NAEP assessment as a whole. These figures illustrate that, across the lowest ability levels, an accessible booklet (i.e., two accessible block) provides a significantly lower estimate of measurement error than either two original NAEP blocks, or the original NAEP assessment as a whole.

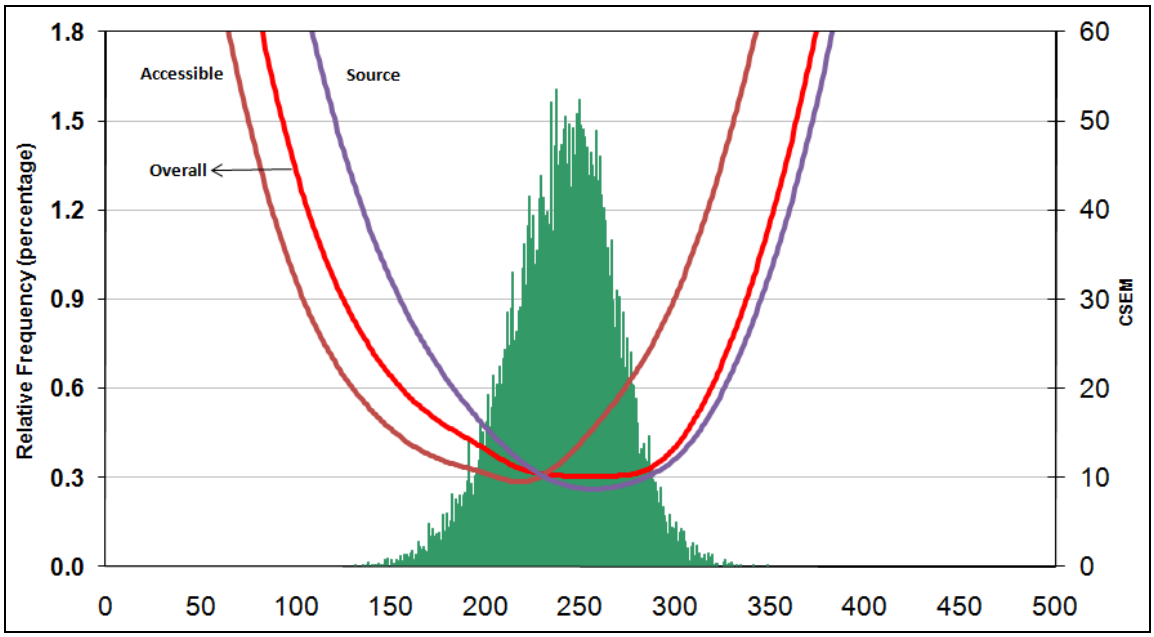


Figure 22. Ability distributions and CSEM by book type – Grade 4.

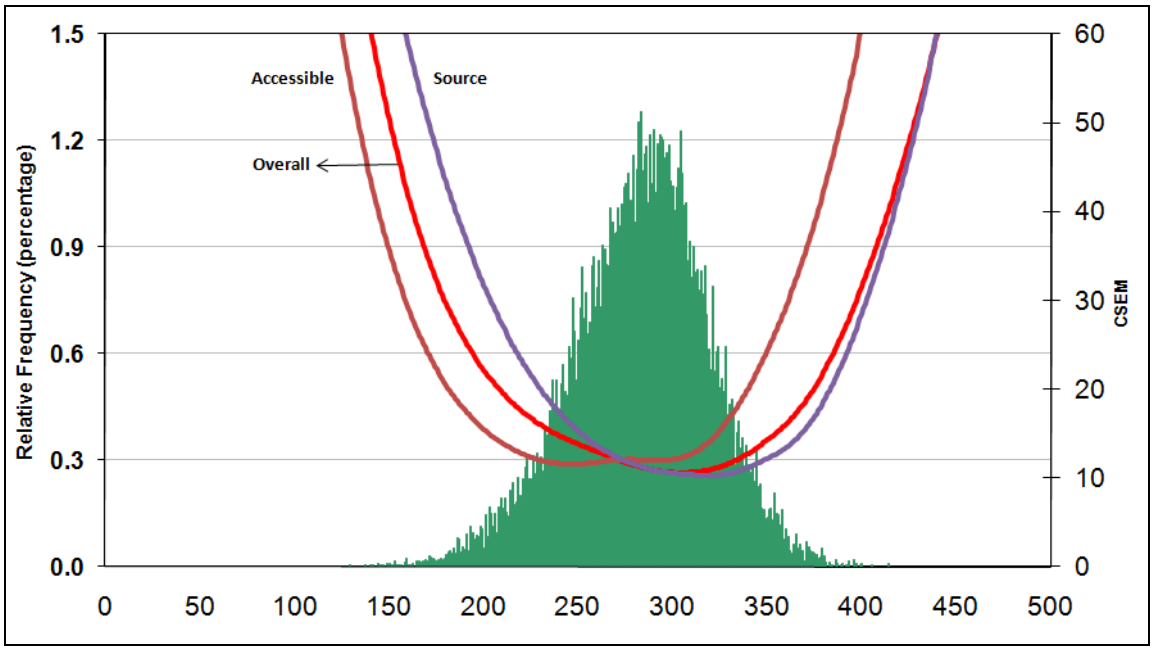


Figure 23. Ability distributions and CSEM by book type – Grade 8

Below, table 20 and table 21 provide point estimates of the reduction in the CSEM across the observed ability distribution for grade 4 and grade 8 students (respectively). These tables show that, for students falling at or below the 25th percentile,

accessible booklets have the potential to provide a significant reduction in observed measurement error, on the order of 20-40 percent.

Table 20

CSEM by Percentile and Book Type – Grade 4

Book Type	5th Percentile	10th Percentile	25th Percentile	50th percentile
Overall	14.2	12.6	10.7	10.2
Accessible	10.8	9.9	9.8	12.6
Source	18.3	15.3	11.6	9.3
Difference	-41%	-35%	-18%	

Note. Difference was computed as 100(source - accessible) / source.

Table 21

CSEM by Percentile and Book Type – Grade 8

Book Type	5th Percentile	10th Percentile	25th Percentile	50th percentile
Overall	17.7	15.2	13.0	11.1
Accessible	12.1	11.6	11.9	12.0
Source	26.3	20.6	14.7	11.6
Difference	-54%	-44%	-19%	

Note. Difference was computed as 100(source - accessible) / source.

Scaling the Accessible Block (RQ3)

Results presented in this section address research question 3: Can accessible items be scaled along with unmodified NAEP items? This section provides a summary of item scaling activities. Item parameter estimates – a, b, and c parameter estimates – for source and accessible blocks are summarized and compared, by grade.

Results indicate that all items in the accessible blocks were scalable with the larger grade 4 and grade 8 item pools NAEP items. Information about students' performance on each item, relative to their estimated proficiency (i.e., theta level), was assessed and item parameter estimates appropriate for, in most cases, the three-parameter logistic model (including item discrimination, difficulty, and guessing parameters) were computed. Accessible items had discrimination and guessing characteristics (a and c parameter estimates) that were generally similar to their source items; while there were significant reductions in item difficulty (b parameter) estimates.

Figure 24 and figure 25 below summarize the observed a parameter estimates for grade 4 and grade 8 students by block type. Figures 26 and 27 below summarize the observed c parameter estimates for grade 4 and grade 8 students by block type. These figures show that, by grade, a and c parameter estimates were generally similar for source and accessible blocks.

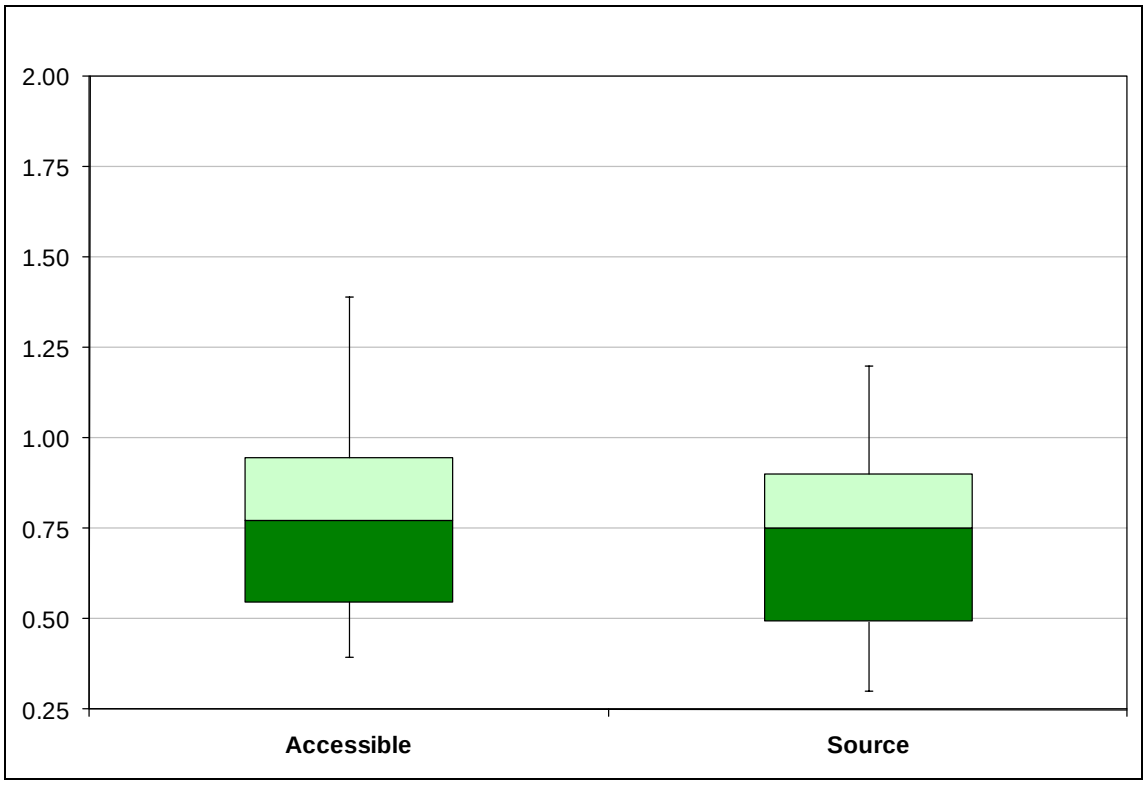


Figure 24. IRT a parameter estimates – Grade 4.

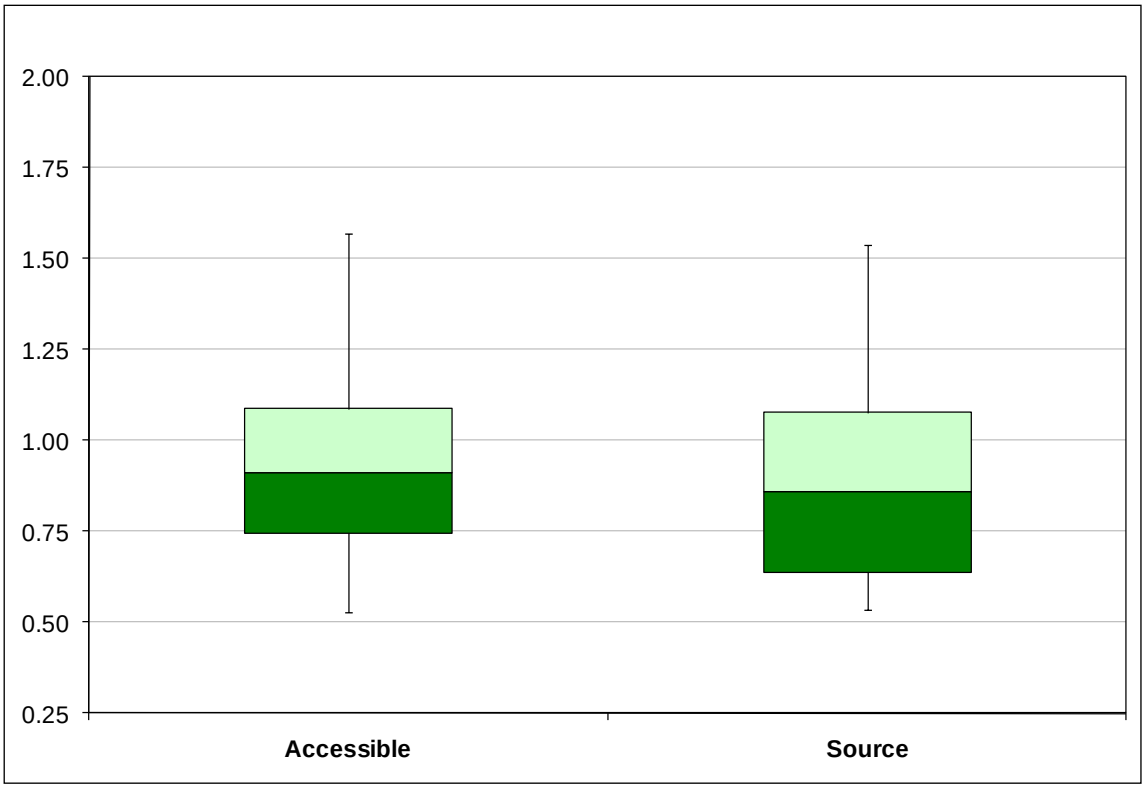


Figure 25. IRT a parameter estimates – Grade 8.

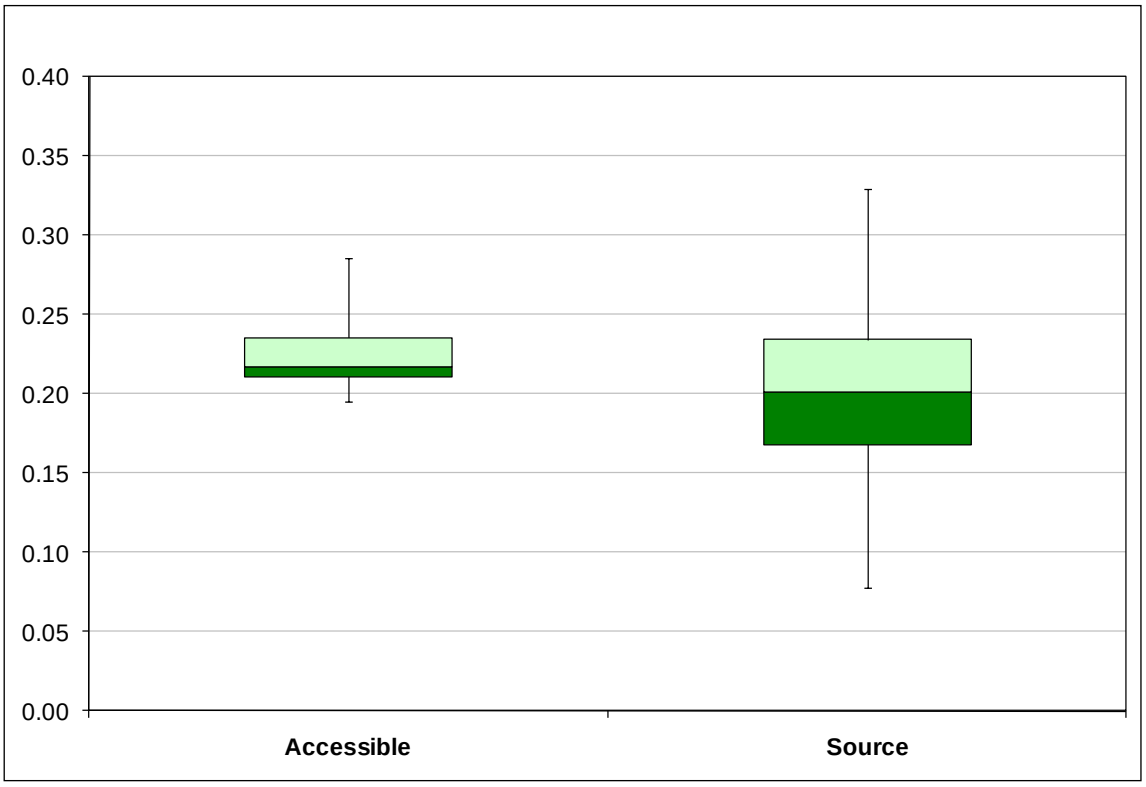


Figure 26. IRT c parameter estimates – Grade 4.

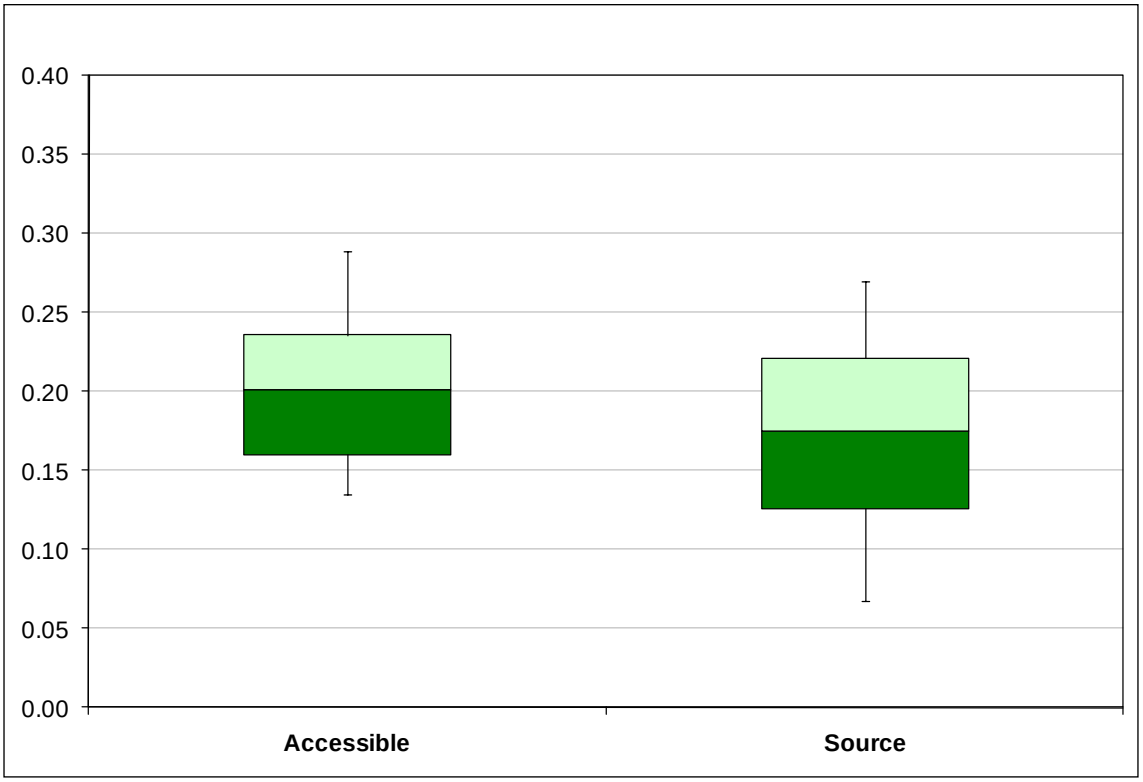


Figure 27. IRT c parameter estimates – Grade 8.

Figures 28 and figure 29 below summarize the estimated b parameter estimates for grade 4 and grade 8 students by block type, and present a summary of the estimated ability distributions for comparison purposes. These figures demonstrate that, on average, accessible items had significantly lower b parameter estimates (item difficulty) than their corresponding source items. That is, for both grade 4 and grade 8, the average estimated difficulty of items included in the accessible blocks was substantially less than the average estimated difficulty of items included in the source blocks. These figures also illustrate that items included in the source blocks were relatively well-aligned with the estimated ability of the sampled student populations, and that items included in the accessible blocks were relatively well-aligned with the estimated ability of students who performed in the lower levels of the NAEP performance continuum.

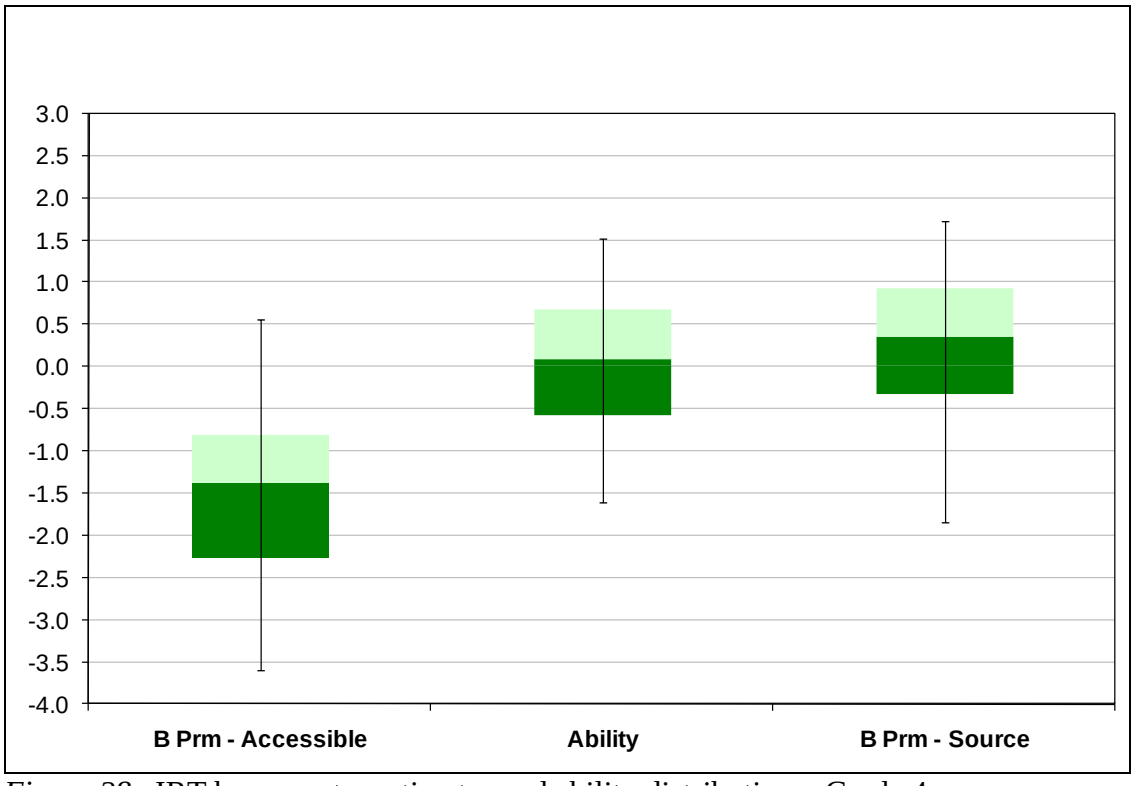


Figure 28. IRT b parameter estimates and ability distribution – Grade 4.

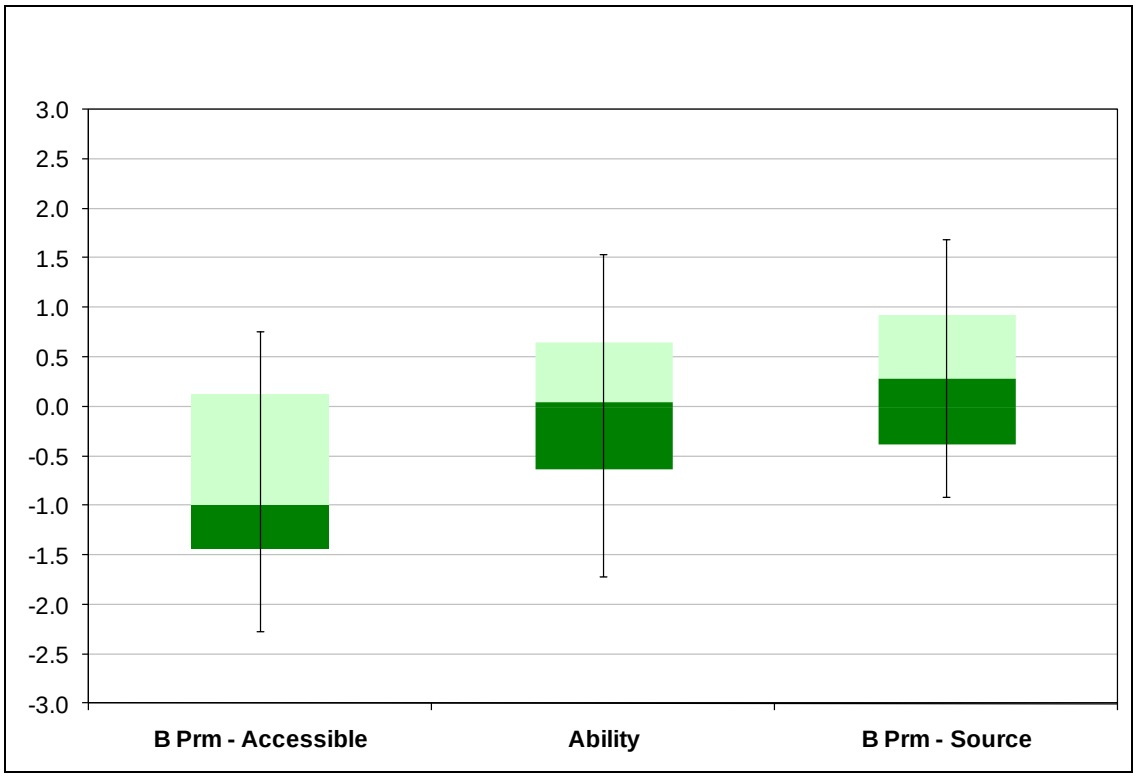


Figure 29. IRT b parameter estimates and ability distribution – Grade 8.

Summary of Results

Across all groups and subgroups there were substantial and similar average gains in percent correct by block. Additionally, there were consistent declines in the number of students omitting various items and significant reductions in the percentage of students not reaching items by grade level. However, item omission and block completion rates varied by disability and English proficiency status. For the lowest performing students, the conditional standard error of measurement was significantly lower on the accessible blocks than the source blocks. Additionally, all items were scalable, and blocks of modified items had similar average discrimination and guessing characteristics (a and c parameter estimates), while significant reductions in item difficulty (b parameter estimates) were observed.

CHAPTER 5

DISCUSSION

This chapter is divided into five major sections. The first section offers a brief overview of the study, summarizing the major objectives, methods, and results that were detailed in the first four chapters. The second section offers a brief discussion which highlights the significance of the study and summarizes the major implications for NAEP and the broader field of educational assessment. The third section summarizes the limitations of this study, and the fourth section offers directions for future research. The chapter closes with a few concluding thoughts about the implications of this study.

Summary of the Study

Precision and reliability are central concerns for any standardized assessment enterprise. Assessments, such as NAEP, that are designed to provide reasonable estimates of student achievement across the performance continuum – and to track trends in student performance over time – are tasked with developing and maintaining assessment structures (e.g., item pools) and procedures (e.g., participant selection protocols) that meet those needs. During recent decades increased attention and resources have been invested in establishing and refining NAEP policies regarding the assessment of subpopulations of students who have historically been marginalized by, or underperformed in, traditional educational systems and structures (e.g., students with disabilities, English language learners). An assumption being made throughout this study – and supported by an evaluation completed by Daro et al. (2007) – was that the quality and accessibility of many of the items in the grade 4 and grade 8 NAEP math item pools was less than ideal. The current study was also initiated under the assumption that the

NAEP assessment could, and perhaps should, more reliably estimate the performance of low-performing students, and endeavored to explore one potential alternative for improving measurement precision (i.e., the accessible block alternative).

The study was conducted in two phases. The first phase of the study focused on the development of a set of *Item Modification Guidelines* and *Item Modification Procedures*, and concluded with a pilot of the accessible block alternative with a relatively small sample of grade 4 students. The second phase of the study focused on applying the *Item Modification Guidelines* and *Item Modification Procedures* to create two accessible blocks at each grade level (grade 4 and grade 8), administering the blocks to nationally representative samples of NAEP participants, and evaluating the results of the study. As detailed in Chapter 3, the process of developing accessible blocks was iterative, and involved multiple activities including expert reviews of original and modified NAEP items, the implementation of the *Item Modification Guidelines* by a group of individuals with diverse expertise in mathematics, assessment, education, and special populations of interest (i.e., students with disabilities and English language learners), cognitive labs, and multiple reviews by NAEP representatives and administrators.

This study explored the feasibility and technical merit of improving the accessibility of existing blocks of grade 4 and grade 8 NAEP math assessment items, and examined changes in student performance and measurement precision that resulted from the implementation of the accessible block alternative. The results presented in Chapter 4 generally indicated that the accessible block alternative significantly increased student performance (i.e., average percent correct), reduced item omission rates, and increased

block completion rates. However, item omission rates and block completion rates varied by disability and English proficiency status. Results also indicated that, for the lowest performing students, estimates of measurement error were significantly lower on the accessible blocks than the source blocks. The sample size obtained for this study did not allow for the estimation of changes in measurement error for other subgroups of interest. The results presented in Chapter 4 further indicated that all grade 4 and grade 8 items included in accessible blocks were scalable with the larger NAEP item pools. Items included in accessible blocks (by grade) had average discrimination and guessing characteristics (a and c parameter estimates) that were similar to items that were included in the source blocks. Also, significant reductions in item difficulty (b parameter estimates) were reported.

Implications

Previous research has found that the error of measurement associated with estimates of student achievement on the grade 4 and grade 8 NAEP math assessments is significantly higher for the lowest performing students (Daro, et al. 2007). Additionally, previous research has suggested that students with disabilities and English language learners are likely to be disproportionately impacted by features of math items that unduly contribute to construct irrelevant variance (Abedi & Hejri, 2004; Mahoney, 2008; Martiniello, 2008). The challenge addressed in this study was to provide NAEP administrators (and the broader community of NAEP stakeholders) with information regarding the feasibility of implementing an accessible block alternative for the grade 4 and grad 8 NAEP mathematics assessments as a means of increasing the precision with which estimates of student achievement, for the lowest performing students, could be

made. This challenge was met by creating and implementing a set of *Item Modification Guidelines* and *Item Modification Procedures* that were intended to provide NAEP administrators – and the broader educational assessment community – with a systematic, empirically based, strategy for improving item accessibility while minimizing or eliminating features of items that contribute to construct irrelevant variance.

Evidence of improved accessibility. The results presented in chapter 4 (and summarized above) provide strong evidence that the average accessibility of items in the accessible blocks was greater than the average accessibility of items in the source blocks. In addition to increased levels of student performance for all groups and subgroups of interest, a reduction in estimated levels of measurement error for the lowest performing students was also observed. These results are logical (fundamental) consequences of increasing item accessibility. That is, as items become more accessible, it is more likely that students who have the requisite mathematics knowledge and skills will correctly answer those items. Increased levels of item accessibility will not increase the likelihood that a student who does not have the requisite knowledge and skills will correctly answer an item. Of course, one would expect similar results when reducing or eliminating item characteristics that contribute to construct irrelevant variance. Here then, it becomes clear that minimizing or eliminating sources of construct irrelevant variance is a reasonable strategy for increasing item accessibility. The application of other item modification strategies, such as adding cues or clarifying alternative answer choices, are also reasonable strategies for improving item accessibility.

The general decreases in item omission rates and increases in block completion rates that were reported in Chapter 4 for the full samples of grade 4 and grade 8 students

participating in this study also indicate that item accessibility was improved. Previous research has identified multiple reasons for item omission including lack of knowledge, missed questions, lack of motivation, lack of time, test-taking strategy, testing conditions, and item format (O'Neil, 1992; Jakwerth & Stancavage, 2003). The link between improved item accessibility, reductions in item omission rates, and increases in block completion rates is, perhaps, less clear than the link between improved item accessibility and increased student performance. However, it is reasonable to assume that increased levels of accessibility would ameliorate several causes of item omission and block incompletion. These sources include lack of knowledge (i.e., construct irrelevant knowledge demands were minimized or eliminated), lack of motivation (i.e., students performed significantly better on the accessible block alternative), and lack of time.

It is important to reiterate that a primary objective of this study was to modify items to increase accessibility while maintaining appropriate alignment with the NAEP frameworks. As Martiniello (2008) notes:

> It is critical that improved accessibility is not achieved at the expense of altering the construct/skill to be measured by the item/test […] Mathematical discourse in classrooms and textbooks combine natural and academic language, mathematical terms, symbols and graphs. So should math assessments, particularly those designed to assess mathematics for understanding (p. 362).

Construct and content validity are central concerns for any standardized assessment, and every effort was made to ensure that high levels of content and construct validity were maintained throughout the item modification process described in this document.

Implementation challenges. The results presented in this study indicate that it may be possible (and appropriate) to incorporate blocks of accessible items as a regular component of the NAEP assessment. That is, the technical characteristics of the items included in the accessible blocks are comparable to the technical characteristics of items included in the full NAEP item pools (i.e., accessible items can be scaled with the full NAEP item pools). However, this study did not endeavor to make claims or judgments regarding current NAEP assessment products, policies, or procedures. It is fully acknowledged that substantial policy questions and technical considerations must be addressed before an accessible block alternative could be fully incorporated into standard NAEP administration practices. Indeed, implementing an accessible block alternative – or any form of two-stage testing – presents significant practical and policy questions for NAEP administrators.

Incorporating the accessible block alternative into the regular NAEP assessment would constitute a significant shift in NAEP policy. One practical and policy related challenge that NAEP administrators would have to address – if they chose to implement and accessible block alternative – would be to evaluate the relative merit of various strategies for incorporating accessible blocks into the full NAEP assessment. Potential criteria for evaluating the relative merit of various strategies for incorporating accessible blocks into the regular NAEP assessment procedures are provided on page 4. Viable strategies for implementing an accessible block alternative may include: (a) offering an accessible booklet as an accommodation to qualifying students, (b) incorporating the accessible blocks into the regular NAEP spiral, or (c) offering the accessible booklet as an accommodation to qualifying students *and* incorporating

accessible blocks into the regular NAEP spiral. Of course, an accessible booklet of items may consist of two blocks of accessible items, or alternatively one block of accessible items could be paired with one block of regular NAEP assessment items to create an accessible booklet. The merit of each of the aforementioned alternatives for incorporating accessible blocks into the full NAEP assessment should be explored.

A second practical and policy related challenge facing NAEP administrators – if they choose to implement an accessible block alternative – is defining an appropriate, politically viable, and empirically based mechanism for identifying students who are the best candidates to complete an accessible block alternative. The results of this study indicate that, for grade 4 and grade 8 students, significant reductions in standard error could be achieved by offering an accessible block to students who fall into the lowest quartile of ability. One possible strategy which NAEP officials could employ to identify the best candidates for accessible block participation would be to review students' performance on previous standardized assessments in order to gauge the likelihood that it would be beneficial to present them with an accessible block. Previous research (Linn, McLaughlin, Jiang, & Gallager, 2004; McLaughlin, Scarloss, Stancavage, & Blankenship, 2005) has indicated that this strategy may be feasible and appropriate.

A second strategy that NAEP officials could employ to identify the best candidates for accessible block participation would be to solicit the professional opinions of school officials (i.e., teachers or counselors) who are most familiar with individual students' current level of mathematics proficiency. Both strategies, however, are complicated by the fact that states, districts, schools, and teachers have varying definitions of student success, and employ various standardized assessments protocols to

evaluate student performance. The degree of divergence in these definitions is large, and reasons for this divergence is not always understood by the public (Linn, 2007). Additionally, states have varying expectations for math learners which may or may not be well aligned with the NAEP frameworks and objectives.

Additional criteria, including SD and ELL status, may also play an important role in identifying students for whom the accessible block is an appropriate assessment alternative. If the accessible block alternative was officially classified as an "accommodation" rather than simply being incorporated into the regular NAEP spiral, then it would be necessary to amend the list of accommodations which are permissible under current NAEP policies to include the accessible block alternative. Many SD and ELL students are currently excluded from NAEP because they are judged (by school officials and NAEP representatives) to be incapable of meaningfully participating in the traditional assessment. Unlike many state assessments, NAEP offers no alternative assessment (which limits NAEP's ability to include these students because alternative assessments are recommended in many IEP and 504 plans). However, research has shown that providing accommodations to students increases their participation rates (Anderson, Jenkins & Miller, 1996; Olson & Goldstein, 1997). It should be noted that NAEP administrators have made strong efforts to investigate viable strategies for increasing the participation of students with disabilities and English language learners – as is evidenced by steadily declining exclusion rates – and the incorporation of the accessible block alternative as a permissible accommodation may further support these efforts.

However, NAEP administrators should carefully consider the ramifications of implementing the accessible booklet option as an accommodation. This study shows that an accessible booklet may serve as an *effective* accommodation (i.e., low performing students perform better on the accessible block and precision is improved), but does not show that the accessible booklet is a *valid* accommodation. Robinson (2010) notes that in order for an accommodation to be valid it should only improve outcomes for the target population(s). If non-target populations benefit from receiving a proposed accommodation, then they should also receive it. Otherwise, the validity of the proposed accommodation may be compromised.

In order for the validity of any accommodation to be assessed, target population(s) must first be identified, and appropriate assessments of validity performed. The results of this study suggest that it may be beneficial to implement an accessible block alternative with a broad range of grade 4 and grade 8 students. Therefore, presenting the accessible booklet as an accommodation to a subsample of low performing students (e.g., SDs or ELLs) may not be viewed as a valid assessment strategy.

Of course, any change made to NAEP assessment protocols, item pools, or participant selection criteria could be viewed as a threat to the validity of the assessment. A primary concern of NAEP administrators would be preserving the ability to perform valid trend analyses (i.e., the ability to measure changes in student performance over time), which have been maintained since 1992 for grade 4 and grade 8 mathematics. Because the incorporation of the accessible block alternative into the full NAEP assessment would constitute a significant modification to the NAEP item pool, it would

be critical for NAEP administrators to complete appropriate technical studies (i.e., linking and equating studies) to ensure the validity of NAEP trend analyses.

Alternative strategies for improving precision. Multiple "low-tech" solutions for reducing the measurement error associated with estimates of achievement for the lowest performing students (or any group of students) are well known. A partial list of possible low-tech solutions for reducing measurement error includes: increasing sample size, oversampling low-performing students, improving the quality of item pools (e.g., increasing discrimination power of items within the pools), and extending test time or length. However, increasing sample sizes and extending testing time or length could be cost prohibitive, and oversampling low-performing students could be viewed as unfair (i.e., placing an inequitable assessment burden on particular subpopulations of students). This suggests that efforts to improve the quality of the NAEP item pool should be pursued. For example, many of the recommendations outlined in the *Item Modification Guidelines* and *Item Modification Procedures* presented in this study could be incorporated into future NAEP item writing activities. As items in the NAEP pool are retired and replaced, such efforts may improve average levels of item accessibility and reduce levels of construct irrelevant variance. The results of this study demonstrate that the systematic application of strategies for increasing accessibility and may have a marked impact on student performance and precision.

Another "low-tech" solution for reducing measurement error would be to present students with one or more blocks of items that, to the greatest degree possible, match their estimated ability levels. Many students find the NAEP assessment difficult, and the majority of items presented to students are designed to best match the expected

performance of students of "average" ability (as illustrated in figure 28 and figure 29). The assessments are not as well suited to the measure the skills and abilities of the lowest performing students, and as a result, the precision with which NAEP officials are able to estimate the achievement of these students is diminished.

By presenting students with items that closely match their ability, it is possible to estimate their achievement with greater reliability. The accessible block alternative presented in this study provides NAEP administrators with an opportunity to do just that. More specifically, the results of item scaling activities which were completed as a part of this study – and reported in Chapter 4 – demonstrate that the average difficulty (i.e., b parameter estimates) associated with items in the accessible blocks are more closely aligned with the estimated ability of the lowest performing students than the NAEP assessment as a whole. Consequently, items in the accessible blocks provide more information about students in this performance range.

However, the accessible block alternative – which could be employed as a component of a two-phase testing strategy – is not ideal. That is, the knowledge and technology exists to implement assessment mechanisms that can more accurately estimate the mathematical abilities of populations of grade 4 and grade 8 students than is possible via two-phase testing. Computer adaptive testing (CAT) is one such solution. Essentially, CAT provides all students (i.e., students of high, average, and low mathematical ability – and gradients thereof) with an opportunity to complete a set of assessment items that closely matches their estimated achievement level. And, a major advantage of CAT is that estimates of student ability can be adjusted based on responses to particular assessment items *during* the assessment process. However, CAT also

requires that all examinees have access a reliable computer terminal with (ideally) a secure internet connection in order to complete the assessment. The infrastructure necessary to administer the NAEP assessment to a nationally representative sample of grade 4 and grade 8 students – which is comparable in size and scope to the current NAEP sample – via computer does not yet exist.

In light of the costs and challenges associated with significantly increasing sample sizes and implementing a computer based version of the assessment, it would seem that some variant of two-phase testing – such as the accessible block alternative – would be a reasonable strategy for increasing precision at the lower levels. Previous research (Bock & Zimowski, 2003; McLaughlin, Scarloss, Stancavage, Blankenship, 2005) indicates that two-stage testing, in particular, has the potential to increase the usability and validity of NAEP results by increasing precision at the lower end of the NAEP performance continuum.

Creating high quality items. Item construction is a central activity in any assessment enterprise. A primary objective for most item development teams is to produce items that will provide a fair and valid assessment of students' skills and abilities. However, determining what is fair and valid is a multifaceted undertaking. For this reason, it is critical that the task of item development not fall solely to content area specialists (Baranowski, 2006; Martiniello, 2008). Such individuals may have a solid understanding of longstanding principles of item writing (e.g., write clearly and concisely), but content specialists alone may be unaware of the consequences of including or excluding particular item features (e.g., graphics), for particular populations (i.e., English language learners), in particular assessment contexts (Abedi, 2006;

Baranowski, 2006). For this reason, it is beneficial to include individuals with multiple competencies in the item writing and review process. In this study, the item writing and modification teams were composed of individuals with mathematics expertise, individuals familiar with issues pertaining to the assessment of students with disabilities and English language learners, current and former educators, as well as NAEP assessment specialists. It may be beneficial to include individuals with similarly diverse sets of competencies and training in other item development contexts.

Similarly, item accessibility is also a central concern for item developers. The application of the *Item Modification Guidelines* and *Item Modification Procedures* created as a part of this study could result in improved accessibility – and reduced levels of construct irrelevant variance – for a broad range of educational assessments. Of course, these documents could be improved through the critical consideration of the broader educational assessment community, as well as the contributions of continued research.

Closing achievement gaps. The persistent achievement gaps associated with populations of students with disabilities, English language learners, and other subgroups of interest are well documented (Lee, 2004; Lubienski, 2002; Robinson, 2010; Robinson & Lubienski, 2011). Additionally, there are continual efforts to improve the skills, abilities, and achievement of these groups. The No Child Left Behind Act of 2001 (i.e., the most recent reauthorization of the Elementary and Secondary Education Act), provides increased levels of accountability and incentive for states, districts, and schools to ensure that all students receive appropriate levels of educational support. Educational achievement is primarily assessed via standardized tests, and the results of these tests

play a key role in shaping the larger educational policy agenda. Naturally, increased attention is paid to populations of students who do not perform well (Linn, 2007).

Educators, administrators, and legislative representatives who champion particular educational reform efforts are asked to demonstrate the impact of the policies and programs which they support. However, gauging the impact of a particular educational reform effort is a difficult task, and progress is often incremental. In this context, then, it becomes increasingly important that standardized measures of student achievement – such as NAEP and state achievement tests – be able to detect incremental changes in student achievement, particularly changes in achievement for the lowest performing students. If some state achievement tests show consistent gains in achievement for the lowest performing students, and NAEP does not, than that may be due to NAEP's inability to detect change with reasonable levels of reliability and precision. That is, the validity of the NAEP assessment rests, in part, in its ability to detect changes in student achievement at the lower end of the performance continuum with relatively high levels of precision. Because the results of this study indicate that accessible blocks reduce the error associated with estimates of student achievement (i.e., increase precision) for the lowest performing students, it should be considered a viable strategy for improving the validity of the results which NAEP administrators report for those students.

Limitations

Ideally, the item development activities performed during phase I and phase II of this study would have been supplemented with small group tryouts with approximately 30 students per test book. These group tryouts could have been used to improve

estimates of item difficulty in the newly constructed blocks. Not only do small group tryouts more closely mimic the demand characteristics of a regular NAEP administration, but the sample sizes, while small, would have represented a significant increase over the 4-8 students per block used in the cognitive labs. It have been beneficial to refine estimates of item difficulty through small group tryouts because only a limited number of blocks could be – and can ever be – advanced for pilot testing and full administration. Small group tryouts would have aided in maximizing the likelihood that selected blocks would succeed. In order to avoid exposure of live NAEP blocks, one could rely on released NAEP blocks to estimate theta levels for the group tryout samples. However, a short timeline prohibited the research team from implementing small group tryouts.

It would also have been beneficial to convene panels of individuals with expertise in issues relevant to the education and assessment of SDs and ELLs to conduct reviews of source and accessible items. These review activities could have been similar in purpose and scope to the math content reviews that were conducted as a part of this study. Although the phase I and phase II item modification panels did bring some level of expertise issues relevant to SDs and ELLs, it was not their sole task to evaluate the appropriateness of source and accessible items for these particular populations. Convening panels of individuals for the sole purpose of evaluating the accessibility of items for SDs and ELLs would have enhanced the fidelity with which the Item Modification Guidelines were applied, and subsequently – the relative accessibility of the items for these particular populations of students. However, it should be noted that the purpose of this study was not to design accessible blocks specifically for implementation for SDs and ELLs. Rather, the purpose of this study was to investigate strategies for

increasing the overall accessibility of the grade 4 and grade 8 NAEP math assessments for low performing students (i.e., a broadly defined portion of the students which NAEP endeavors to assess).

Directions for Future Research

The NAEP item pools regularly undergo examination from multiple entities, and efforts to improve item quality and reduce construct irrelevant variance should continue. The expert review activities conducted as a part of this study, and supported by a recent evaluation of the grade 4 and grade 8 NAEP math item pools (Daro, Stancavage, Ortega, DeStefano, & Linn, 2007), suggest that the quality of the grade 4 and grade 8 NAEP mathematics item pools is less than ideal. Many items have minor or significant flaws (i.e., diminished mathematical validity or sources of construct irrelevant variance) that could be minimized or eliminated. The *Item Modification Guidelines* and *Item Modification Procedures*, which were created as a part of this study and based in part on existing NAEP item writing guidelines, may provide some guidance in this regard. Of course efforts to improve the quality of the grade 4 and grade 8 NAEP item pools would also contribute to efforts to bolster the construct validity of the assessment.

Although the results of this study are promising, more research is needed to determine the extent to which the incorporation of accessible booklets into the regular administration of NAEP increases precision at the lower end of the performance continuum. Additionally, efforts should be made to investigate how the incorporation of accessible booklets improves measurement precision for low-performing students who are also English language learners or students with disabilities. Because the sample of English language learners and students with disabilities who participated in this study

was relatively small, it was not feasible to complete such analyses as a part of the current research effort. Future investigations should include larger numbers of these students, and their performance on accessible blocks of items should be thoughtfully evaluated.

Future research could focus on improving our understanding of how various types of item modifications impact on student performance, and for which subpopulations of students these modifications are most effective. This study did not endeavor to address such questions, but the capacity to do so exists. The item modification guidelines produced during phase I of this study provided item writers with suggestions for modifying particular characteristics of existing NAEP items to make them more accessible. In this study, several item features are discussed (i.e., graphics, language, alternative answer choices, cues, etc.), but little guidance was provided to item writers and reviewers regarding the relative importance of attending to various item features. Complex grammatical structures, for example, may be a source of construct irrelevant variance for some groups of students, but not others. The goal of such research would be to better understand how to ameliorate construct irrelevant variance (i.e., bias) for particular subpopulations of interest.

For example, in this study efforts to improve accessibility appeared to differentially affect students with disabilities and English language learners. Results presented Chapter 4 indicate that SD and ELL students were, on average, able to perform significantly better on the accessible block version of the assessment than the source block version. However, results also indicated that the accessible block alternative did not have a uniformly positive impact on item omission and block completion rates for

these groups. It is important that efforts be made to understand how and why these results occurred.

Similar research has been conducted in the past, and continues to shape our understanding of effective item writing practices today. A recent randomized study conducted by Moreno, Pirritano, Allred, Calvert, & Finch (2006) evaluated the potential value of adding visual representations to text-only math word problems. This research team found that illustrative pictures were an ineffective accommodation for the math assessment of ELLs. Martiniello (2008) notes that this result may be attributed to the fact that most, if not all, of the representations utilized in the cited study were primarily pictorial, and that the potential of using schematic representations combined with text in math word problems should be further examined.

Because the results of this study are promising, NAEP administrators may explore the possibility of developing an accessible block alternative for other content areas that are currently assessed by NAEP, such as reading and science. The item modification guidelines and item modification procedures that were drafted during phase I of this study and refined during phase II of the study offer a clear description of the types of modifications that can be NAEP math items in order to improve item accessibility and reduce construct irrelevant variance, and it may be possible to adapt these guidelines for use with other subject area assessments. It is not clear if it would be necessary to make significant additions or modifications to the item modification guidelines and procedures in order to meaningfully apply the item to assessments in other content areas, but such modifications should be possible.

Although the results of this study are promising, NAEP administrators should carefully examine the implications of implementing an accessible block alternative (e.g., validity concerns, analysis concerns, costs) as a means of improving measurement precision. Similarly, NAEP administrators should examine the viability of various strategies for implementing the accessible block alternative with various subpopulations of interest.

Conclusion

The NAEP provides a context and platform for meaningful conversations regarding national education policy, and serves as a fulcrum for educational reform. Multiple constituencies leverage NAEP results to support (or contradict) important claims about educational systems, styles, and approaches. It is the validity of such claims that must be subjected to constant scrutiny (Della-Piana, 2008). Legislation, including Race to the Top (i.e., the major educational component of the American Recovery and Reinvestment Act of 2009), funnels billions of dollars from federal and state governing bodies to various districts and educational agencies based, in part, on NAEP results. Clearly, the stakes are high. It is important that NAEP results be as accurate as possible, particularly for those students who have historically been marginalized by the educational system (e.g., students with disabilities, English language learners, low-achieving students), and are so often the subject of educational policy debates and the target of various intervention programs and reforms.

The NAEP Validity Studies Panel and other groups have been interested in the use of modified blocks as a means of improving measurement at the lower levels of the NAEP scale and increasing the accessibility and validity of NAEP for all students. The

aim of including one or more accessible blocks would not be to make NAEP easier, but to improve measurement at the lower end of the performance continuum by including more items that provide information about those students' abilities and skills. Increased precision at the lower levels represents an important validity issue regarding the use of NAEP as a means of benchmarking and interpreting both status and change in student achievement over time.

Summary

This study builds on existing efforts to understand strategies for – and the consequences of – increasing item accessibility. The findings presented here suggest that increased levels of accessibility can have a significant impact on student performance and enhance the precision with which the achievement of the lowest performing students is measured. There is still much to learn. For example, can accessible blocks be implemented as a valid accommodation for students with varying degrees of ability, or disability, or English proficiency? If so, what item modification strategies are most effective (i.e., result in increased levels of accessibility), and for whom? This study did not attempt to address such questions. Opportunities to investigate the means and consequences of increasing item accessibility, particularly for subpopulations of interest, are abundant.

REFERENCES

Abedi, J. (2003, April). *Impact of linguistic factors in content-based assessment for ELL students: An overview of research*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Abedi, J., Hejri, F. (2004). Accommodations for students with limited English proficiency in the National Assessment of Educational Progress. *Applied Measurement in Education*, *17*(4), 371-392.

Aiken, L.R. (1972). Language factors in learning mathematics. *Review of Education Research,* 42, 359-385.

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Waveland Press.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Anderson, N., Jenkins, F., & Miller, K. (1996). *NAEP inclusion criteria and testing accommodations: Findings from the NAEP 1995 field test in mathematics.* Princeton, NJ: Educational Testing Service.

August, D., & Hakuta, K. (Eds.). (1998). *Educating language minority children*. Washington, DC: National Academy Press.

Baker, E. (1995). Introduction: Policy and technical contexts of National Assessment of Educational Progress validity studies. *Educational Assessment, 3*(1), 1-8. Retrieved from Academic Search Premier database.

Beaton, A.E., & Zwick, R. (1992). Overview of the National Assessment of Educational Progress. *Journal of Educational Statistics, 17*(2), 95-109.

Berends, M., & Koretz, D. (1995). Reporting Minority Students' Test Scores: How Well Can the National Assessment of Educational Progress Account for Differences in Social Context? *Educational Assessment, 3*(3), 249-289. Retrieved from Academic Search Premier database.

Bohrnstedt, G. W., & Stancavage, F. B., (2007). *Estimating effects of non-participation on state NAEP scores using empirical methods*. NAEP Validity Studies Panel: National Center for Education Statistics.

Bracey, G. (2009). Big tests: What ends do they serve? *Educational Leadership, 67*(3), 32-37. Retrieved from Academic Search Premier database.

Braswell, J. S., Lutkus, A. D., Grigg, W.S., Santapau, S. L., Tay-Lim, B. S. H., & Johnson, M. S. (2001) *The nation's report card: Mathematics 2000*. (NCES 2001-517). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

Bock, D. R., & Zimowski, M. F. (2003). *Feasibility studies of two-stage testing in large scale educational assessment: Implications for NAEP*. NAEP Validity Studies Working Paper Series: National Center for Educational Statistics.

Buckendahl, C. W., Plake, B. S., & Davis, S. L. (2009a). Conducting a lifecycle audit of the National Assessment of Educational Progress. *Applied Measurement in Education, 22*(4), 321-338. doi:10.1080/08957340903221642.

Buckendahl, C. W., Plake, B. S., & Davis, S. L. (2009b). *Evaluation of the National Assessment of Educational Progress*. U.S. Department of Education, Buros Institute for Assessment Consultation and Outreach.

Chromy, J. R. (2003). *NAEP validity studies: The effects of finite sampling corrections on State Assessment sample requirements.* (NCES 2003-17). Washington, DC: U.S. Department of Education, office of Educational Research and Improvement, National Center for Education Statistics.

Cizek, (2001). Conjectures on the rise and call of standard setting: An introduction to context and practices. In G. J., Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives*, (pp. 3-17). Hillsdale, NJ: Erlbaum.

Cocking, R., & Mestre, J. (Eds.) (1988). *Linguistic and cultural influences on learning mathematics*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Creswell, J. W., & Planko Clark, V.L. (2007). *Designing and Conducting Mixed Methods Research*. Thousand Oaks, CA: Sage Publications, Inc.

Cronbach, L. (1980). Validity on parol: How can we go straight? *New Directions for Testing and Measurement, 5*, 99-108.

Daro, P., Stancavage, F. B., Ortega, M., DeStefano, L., Linn, R. (2007). *Validity study of the NAEP mathematics assessment: Grades 4 and 8*. NAEP Validity Studies Working Paper Series: National Center for Educational Statistics.

Della-Piana, G. M. (2008, April). Enduring issues in educational assessment. *Phi Delta Kappan, 89*(8), 590-592.

Haertel, E. H. (2003). *Including students with disabilities and English language learners in NAEP: Effects of differential inclusion rates on accuracy and interpretability of findings*. ED 500430 Retrieved from http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/3d/0e/b6.pdf.

Hambleton, R. K, Brennan, R. L., Brown, W., Dodd, B. Forsythe, R. A., Mehrens, W. A., et al. (2000). A response to "setting reasonable and useful performance standards: in the National Academy of Sciences "Grading the Nation's Report Card." *Educational Measurement: Issues and Practice, 19*(2), 5-14.

Houser, J. (1995). Assessing students with disabilities and limited English proficiency. Working paper series. National Center for Educational Statistics. Washington, D.C.

Jaeger, R. M. (2003). NAEP validity studies: Reporting the results of the National Assessment of Educational Progress (Working Paper 2003-11). Washington, DC: U.S. Department of Education, Institute of Education Services.

Jakwerth, P. M., Stancavage, F. B., (2003). An investigation of why students do not respond to questions. NAEP Validity Studies. Working Paper Series (Report No. NCES-WP-2003-12). Retrieved from http://nces.ed.gov/pubs2003/200312.pdf.

Johnson, E. G. (1992). The design of the National Assessment of Educational Progress. *Journal of Educational Measurement, 29*(2), 95-110.

Kane, M. T. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64*, 425-461.

Koretz, D., Lewis, E., Skewes-Cox, T, & Burnstein, L. (1993). *Omitted and not-reached items in mathematics in the 1990 National Assessment of Educational Progress (CRE Technical Report 347)*. Los Angeles, CA: Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Kipplinger, V., Haug, C., & Abedi, J. (2000, April). *Measuring math – not reading – on math assessment: A language accommodations study of English language learners and other special populations*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Lane, S., Zumbo, B., Abedi, J., Benson, J., Dossey, J., Elliott, S., et al. (2009). Prologue: An Introduction to the Evaluation of NAEP. *Applied Measurement in Education, 22*(4), 309-316. doi:10.1080/08957340903221436.

Lee, J. (2004). Multiple facets of inequity in racial and ethnic achievement gaps. *Peabody Journal of Education, 79*(2), 51-73.

Linn, R. (2006). Following the standards: Is it time for another revision? *Educational Measurement: Issues and Practice*, *25*(3), 54-56.

Linn, R. (2007). Validity inferences from test-based educational accountability systems. *Journal of Personnel Evaluation in Education,* 19(1), 5-15.

Linn, R. L., Koretz, D. M., Baker, E. L., & Burstein, L. (1991). *The validity and credibility of the achievement levels for the 1990 National Assessment of Educational Progress in mathematics*. (Technical Report). Los Angeles: University of California, Center for the Study of Evaluation.

Linn, R., McLaughlin, D.H., Jiang, & Gallagher (2004). *Assigning adaptive NAEP booklets based on state assessment scores: A simulation study of the impact on standard errors*. NAEP Validity Studies Paper Series: National Center for Educational Statistics.

Loomis, S. C. (2001). *Judging evidence of the validity of the National Assessment of Educational Progress achievement levels*. National Assessment Governing Board. ED 459212. Washington D.C. Retrieved from http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/19/7e/6f.pdf.

Lubienski, S. (2002). A closer look at Black-White mathematics gaps: Intersections of race and SES in NAEP achievement and instructional practices data. *The Journal of Negro Education, 71*(4), 269-287.

Mahoney, K. (2008). Linguistic influences on differential item functioning for second language learners on the National Assessment of Educational Progress. *International Journal of Testing, 8*(1), 14-33.

Martiniello, M. (2008). Language and the performance of English-language learners in math word problems. *Harvard Educational Review*, 78(2), 333-368.

McLaughlin, D.H. (2000, July). *Protecting state NAEP trends from changes in SD/LEP inclusion rates*. Paper presented at the National Institute of Statistical Sciences workshop on NAEP inclusion strategies. Research Triangle Park, NC, July 2000.

McLaughlin, D.H. (2001, November). *Exclusions and accommodations affect state NAEP gain statistics: Mathematics, 1996 to 2000*. Report to the NAEP Validity Studies Panel. Palo Alto, CA: American Institutes for Research.

McLaughlin, D.H., Scarloss, B.A., Stancavage, F.B., Blankenship, C.D. (2005). *Using state assessments to assign booklets to NAEP students to minimize measurement error: An empirical study in four states.* A publication of the NAEP Validity Studies Panel. Palo Alto, CA: American Institutes for Research.

Messick, S. (1990). *Validity of test interpretation and use*. Research Report RR-90-11. Princeton, NJ: Educational Testing Service.

Munro, (1979). Language abilities and maths performance. *Reading Teacher, 32*(8), 900-915.

National Assessment of Educational Progress Authorization Act, Pub. L. No. 107-279, 20 U.S.C. § 9622 (2002).

National Assessment Governing Board. (2004). *Mathematics Framework for the 2005 National Assessment of Educational Progress*. Washington, DC: Author.

National Assessment Governing Board. (2006). *Reporting, release, and dissemination of NAEP results: Policy Statement*. Washington, DC: Author.

National Center for Education Statistics. (2010). *The nation's report card: Mathematics 2009* (NCES 2010-451). Washington DC: U.S Department of Education.

National Center for Education Statistics. (2003). *NCES handbook of survey methods* (NCES 2003-603). Washington DC: U.S Department of Education.

National Research Council. (1999). *High stakes: Testing for tracking, promotion, and graduation*. Washington, DC: National Academy Press.

Noell, J., & Ginsburg, A. (2009). Evaluation of the National Assessment of Educational Progress: Next Steps. *Applied Measurement in Education, 22*(4), 409-414. doi:10.1080/08957340903221691.

No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107-110, 115 U.S.C. § 1425 (2002).

Olson, J., & Goldstein, A. (1997). *The inclusion of students with disabilities and limited English proficient students: A summary of recent progress*. Washington, DC: U.S. Department of Education, National Center for Educational Statistics.

O'Neil, H.F. (1992). *Experimental studies on motivation and NAEP test performance*. Final Report. National Center for Research on Evaluation.

Orr, E. (1987). *Twice as less: Black English and the performance of black students in mathematics and science*. New York: Norton.

Pellegrino, J. (2007). Should NAEP performance standards be used for setting standards for state assessments? *Phi Delta Kappan, 88*(7), 539-541. Retrieved from Academic Search Premier database.

Pomplun, M. R. & Omar, M. H. (2001). Factorial invariance of a test of reading comprehension across groups of limited English proficiency students. *Applied Measurement in Education, 14*(3), 261-284.

Qian, J., Kaplan, B. A., Johnson, E. G., Krenzke, T., & Rust, K. F. (2001). Weighting procedures and estimation of sampling variance for the national assessment. In N. L. Allen, J. R. Nonoghue, & T. L. Schoeps, *The NAEP 1998 Technical Report* (NCES, 2001-509). Washington, DC: National Center for Education Statistics.

Resnick, L. (1998). *Reflections on the future of NAEP: Instrument for monitoring or for accountability?* (Report No. CSE-TR-499).

Rivera, C., & Collum, E. (Eds.). (2006). *State assessment policy and practice for English language learners*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Robinson, J. P. (2010). The effects of test translation on young English learners' mathematics performance. *Educational Researcher, 39*(8), 582-590.

Robinson, J. P., & Lubienski, S. (2011). The development of gender achievement gaps in mathematics and reading during elementary and middle school: Examining direct cognitive assessments and teacher ratings. *American Educational Research Journal, 48*(2), 268-302.

Rothman, R.W., & Cohen, J. (1989). The language of math needs to be taught. *Academic Therapy, 25*, 133-142.

Rothman, R., Slattery, J.B., Vranek, J.L., & Resnick, L.B. (2002). *Benchmarking and alignment of standards and testing*. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing. CSE Technical Report 566.

Rust, K. F., & Johnson, E. G. (1992). Sampling and weighting in the national assessment. Journal of Educational Statistics, *17*(2), 111-129.

Sireci, S. G., Hauger, J. B., Wells, C. S. Shea, C., & Zenisky, A. L. (2009). Evaluation of the standards setting on the 2005 Grade 12 National Assessment of educational Progress Mathematics Test. *Applied Measurement in Education, 22*(4), 409-414. doi: 10.1080/08957340903221659.

Stern, L., & Ahlgen, A. (2002). Analysis of students' assessments in middle school curriculum materials: Aiming precisely at benchmarks and standards. *Journal of Research in Science Teaching, 39*(9), 889-910.

Stoneberg, B. (2005). Please don't use NAEP scores to rank order the 50 states. *Practical Assessment, Research & Evaluation, 10*(9). Retrieved from http://pareonline.net/getvn.asp?v=10&n=9.

Stufflebeam, D., Jaeger, R. M., & Scriven, M. (1991, August). *Summative evaluation of the National Assessment Governing Board's inaugural 1990-91 effort to set achievement levels on the National Assessment of Educational Progress*. Washington, DC: National Assessment Governing Board.

Swinton, S. (1991). *Differential response rates to open-ended and multiple-choice NAEP items by ethnic groups*. Unpublished manuscript, Oct. 23. Princeton, NJ: Educational Testing Service.

Tashakkori, A., & Teddlie, C. (1998). *Mixed Methodology: Combining Qualitative and Quantitative Approaches*. Thousand Oaks, CA: Sage Publications, Inc.

Tyler, R. (1966). New trends in education. *American Journal of Psychiatry, 122*(12), 1394-1398.

Vinovskis, M. A. (1998). *Overseeing the nation's report card: the Creation and evolution of the National Assessment Governing Board*. Washington, DC: National Assessment Governing Board.

Webb, N. L., (1997). Research monograph No. 6: Criteria for alignment of expectations and assessment in mathematics and science education. Washington, DC: Council of Chief State School Officers.

Wells, C., Baldwin, S., Hambleton, R., Sireci, S., Karatonis, A., & Jirka, S. (2009). Evaluating score equity assessment for state NAEP. *Applied Measurement in Education, 22*(4), 394-408. doi:10.1080/08957340903221683.

Williams, V. (1999). *Exploring statistical adjustment of results from the national assessment of educational progress*. National Institute of Statistical Sciences.

Zieky, M. (2001). So much has changed: How the setting of cutscores has evolved since the 1980s. In *Setting Performance Standards: Concepts, Methods, and Perspectives*. Edited by G. Cizek. Mahwah, N.J.: Lawrence Erlbaum pp. 19-51.

Zenisky, A., Hambleton, R., & Sireci, S. (2009). Getting the message out: An evaluation of NAEP score reporting practices with implications for disseminating test results. *Applied Measurement in Education, 22*(4), 359-375. doi:10.1080/08957340903221667.

Zhu, D., & Thompson, T. D. (1995). *Gender and ethnic differences in tendencies to omit responses to multiple-choice tests using number-right scoring*. (Eric Document Reproduction Service No. ED 382 689).

Zirkel, P.A. (1972). Spanish-speaking students and standardized tests. *Urban Review, 5*, 32-40.

APPENDIX A

ACCESSIBILITY

This section is drawn from the 2005 NAEP Mathematics Assessment and Item Specifications. This information has been recreated and included here because accessibility is a key consideration when creating accessible blocks.

Accessibility

Accessibility in an educational assessment context refers to the degree to which the assessment provides all students in the targeted population with the opportunity to demonstrate their achievement in relation to the construct of interest. In this case the target population includes SD and ELL students and the construction of interest is students' mathematics achievement on learning objectives that are defined by the NAEP framework. The design for the NAEP mathematics accessible block must address issues of student accessibility—considerations that can either facilitate or block the goal of obtaining valid measurements of the targeted test takers' achievement in mathematics.

The NAEP mathematics assessment is designed to measure the achievement of students across the nation. Therefore, it should allow students who have learned mathematics in a variety of ways, following different curricula and using different instructional materials; students who have mastered the content to varying degrees; students with disabilities; and students who are English-language learners to demonstrate their content knowledge and skill. T he question to ask in developing the assessment is "What is a reasonable way to measure the same intended constructs for students who come to the assessment with different experiences, strengths, and challenges, who approach the constructs from different perspectives, and who have different ways of displaying their knowledge and skill?"

The central requirement of such an assessment is that the same mathematical constructs are measured across diverse groups of students. To this end, the assessment should maintain the rigor of the mathematics expectations in the framework while providing the means for all tested students to demonstrate their levels of knowledge and skills.

Two methods NAEP uses to design an accessible assessment program are 1) developing the standard assessment so that it is accessible; and, 2) providing accommodations for students with special needs.

Test Accessibility Components

Multiple access points appropriate for the diverse population of students should be available throughout the assessment. Ways to strengthen access include the following:

1. Paying careful attention to how items are presented to students in the assessment (e.g., plain language and editing procedures, use of graphics, item format considerations, use of manipulatives or other tools).
2. Designing constructed-response items so that they allow for multiple ways of responding, as appropriate to the knowledge and skill assessed.
3. Developing scoring rubrics so that the targeted knowledge and skills are evaluated at all score levels.
4. Formatting assessment booklets to allow enough space between items; using boxes and lines judiciously.

APPENDIX B

CONTENT EXPERT PANEL MEMBERS

Phase I Content Expert Panel Members

Patrick Callahan
University of California, Office of the President
email: Patrick.Callahan@ucop.edu

Arthur (Art) Duval
University of Texas, El Paso
email: artduval@math.utep.edu

Roger Howe
Yale University
email: howe@math.yale.edu

Wilfried Schmid
Department of Mathematics
Harvard University
email: schmid@math.harvard.edu

Phase II Content Expert Panel Members

Peter Braumfield
University of Illinois at Urbana Champaign
email: pbraunfe@illinois.edu

Patrick Callahan
University of California, Office of the President
email: Patrick.Callahan@ucop.edu

Arthur (Art) Duval
University of Texas
email: artduval@math.utep.edu

Roger Howe
Yale University
email: howe@math.yale.edu

Randy McCarthy
University of Illinois at Urbana Champaign
email: rmccrthy@illinois.edu

Wilfried Schmid
Harvard University
email: schmid@math.harvard.edu

APPENDIX C

ITEM MODIFICATION PANEL MEMBERS

Phase I Item Modification Panel Members

Panel Coordinator

Jeremiah Johnson
email: jeremiahmatthewjohnson@yahoo.com

Phase I Panel Members

Hsin-Mei Huang, Ph.D.
email: hhuang22@illinois.edu

Renee Lemons
email: rlemons@illinois.edu

Travis Wilson
email: wilson2@illinois.edu

Phase II Item Modification Panel Members

Panel Coordinator

Jeremiah Johnson
email: jeremiahmatthewjohnson@yahoo.com

Panel Members

Holly Downs
email: hadowns@illinois.edu

Kathleen R. Smith
email: smithka@illinois.edu

Aaron Hill
email: aaronthill@gmail.com

Jason Pound
email: jpound@usd116.org

Renee Lemons
email: rlemons@illinois.edu

Theresa Bryant
email: tbryant@usd116.org

Guy Tal
email: guytal2@uiuc.edu

Tony Se
email: tonyse@illinois.edu

Jacqueline Bunn
email: jbunn1@illinois.edu

APPENDIX D

ITEM MODIFICATION GUIDELINES AND PROCEDURES

Aligning the Accessible Block Assessment with the NAEP Framework

Similar to standard blocks of NAEP assessment items, all accessible blocks should be developed so that they are aligned with the content expectations defined by the 2005 NAEP Mathematics Framework. Unlike standard blocks of NAEP assessment items, there will be less variability in the level of complexity of items in a NAEP accessible block. Drawing upon Webb and others*, five interrelated dimensions are considered in structuring the NAEP assessment so that it is aligned with the NAEP framework:

1. The match between the content of the assessment and the content of the framework: The assessment as a whole should reflect the breadth of knowledge and skills covered by the topics and objectives in the framework.

2. The match between the complexity of mathematical knowledge and skills on the assessment and in the framework: The assessment as a whole should represent the balance of levels of mathematical complexity at each grade level as described in the framework. However, an accessible block is meant to provide important statistical information about students at the lower end of the performance continuum. Therefore, it is appropriate for an accessible block to contain items that assess students' ability to perform tasks associated with Basic and Proficient levels of achievement.

3. The match between the emphasis of the assessment and the emphasis of topics, objectives, and contextual requirements in the framework: The assessment should represent the balance of content and item formats specified in the framework and give appropriate emphasis to the conditions in which students are expected to demonstrate their mathematics achievement, reflecting the use of calculators, manipulatives, and real-world settings.

4. The match between the assessment and how scores are reported and interpreted: The assessment should be developed so that scores will reflect both the framework and the performance described in the NAEP achievement levels.

5. The match between the assessment design and the characteristics of the targeted assessment population: The assessment should give all students tested a reasonable opportunity to demonstrate their knowledge and skills in the topics and objectives covered by the framework (with a special emphasis here being placed on providing students at the lower end of the performance continuum an opportunity to show what they know and are able to do).

Item Modification Guidelines

These guidelines identify the major themes and dimensions of construction that should be addressed when modifying blocks of NAEP items to create an accessible block. The guidelines are meant to aid item modifiers in assessing relevant and irrelevant aspects of the item's construct that contribute to the overall difficulty and accessibility of the item. The guidelines offer common strategies for reducing difficulty without compromising content, construct validity or alignment with the NAEP framework.

These guidelines should be applied judiciously. Their application may vary from item to item depending upon the measurement intent of the item. Generally, these guidelines should be followed unless the construct targeted by an item precludes doing so.

Word Choice

Careful word choice is an essential component of quality item construction. Word choice refers to language used within the statement of a problem, as well language used in the alternative answer choices. Careful word choice should be a central consideration during the item modification process.

- **Clarity** - Word choice throughout all items should be unambiguous and concise. It is more important for item wording to be clear than for it to be precise. For example, avoid the ambiguous phrase "about how much" when writing problems that require estimation or rounding.
- **Plain Language** - Plain language, as a writing and editing tool, is designed to clearly convey meaning without altering what an item is intended to measure. All items should use plain language. Even when the intent of the item is for the student to define, recognize, or use mathematics vocabulary correctly, the surrounding text should be in plain language. Plain language should increase access and minimize confusion.
- **Terminology Appropriateness** - Terminology used should be current and relevant to a broad population. Use of outdated technology, terminology, etc. can distract from the content of a problem.
- **ESL Considerations** - Use of commonly accepted and culturally non-specific words, phrases, and terminology is encouraged whenever possible. Be careful of literal interpretations of items. When using words with multiple meanings, make sure the intended meaning is clear. Avoid ambiguous words such as if, could, may, can, etc. Use high-frequency words as much as possible. Avoid the word "not" whenever possible.
- **Parallel Item Construction** - Item wording should provide parallel syntactic construction. Use of the present tense verb is preferred. Wording within and between the statement of a problem and its possible answer choices (including distracters) should be consistent in tense and vocabulary.
- **Brevity and Simplicity** - Questions should have brief, 'simple' form. Compound sentences should be written as two short sentences.

- **Grammar** - Present tense and active voice should be used whenever possible. Minimize paraphrasing. Avoid pronouns. Avoid colloquialisms.

Alternative Answer Choices (Distracters)

Alternative answer choices include the solutions presented in a multiple choice item, as well as the acceptable answers for an open ended item. Alternative answer choices may be presented in multiple formats (numbers, text, graphics, charts, etc.). Use of these formats can increase item access. However, if used or constructed improperly, they can add confusion to the item and may distract test takers from the original intent of the item.

For Multiple Choice Items:

- **Provide Plausible Distracters** - Identify alternative answer choices (distracters) that are plausible, and not unreasonable. The easiest multiple choice questions should provide students with only one reasonably appropriate solution.
- **Provide an Appropriate Number of Distracters** - Make the number of possible answer choices appropriate for the content and context of the problem. The American convention of providing four answer choices is sometimes inappropriate or unreasonable.
- **Provide a Range of Distracters** - Provide students with a diverse set of answer choices. This may reduce confusion and testing error. Items requiring rounding or estimation are sometimes clearer when a wide range of answer choices is provided.

For Open Ended Items:

- **Allow for Multiple Response Types** - Allow students to show their answers through illustrations, diagrams, formulas, or words.

Item and Block Format

Item and block format is the layout, design, and arrangement of information within and between each item in a block. Careful item and block formatting can improve the clarity of an item, and of the block as a whole.

- **Format Consistency** - Use the same structure for paragraphs throughout the assessment as much as possible (e.g., topic sentence, supporting sentences, and concluding sentence). Be sure that item format does not add ambiguity to the solution.
- **Separate Information as Appropriate** - Split multiple ideas into separate sentences or statements, or even separate lines to decrease the complexity of an item.
- **Item Spacing** - Provide liberal spacing throughout an item. Double spacing makes word problems easier to read and understand. Double spacing alternative answer choices aids in visual and cognitive processing and discrimination. Separate the main question in an item (how, what…) from the rest of the information presented in the item.
- **Answer Spacing** - Provide appropriate space for an answer. Too little or too much space for an answer can falsely suggest an answer of a certain length.

- **Clarity** - Use format to clarify text. Use bullets, space between pieces of text, and boxing of text to emphasize or separate information.
- **Item Separation** - Provide a clear distinction between each item. Some NAEP items provide information (e.g., a graph, chart) before the statement of the problem. In such cases, the item should always begin on a new page in order to provide a clear distinction between problems.

Graphics

Graphics such as pictures, charts, and diagrams are visual images reflecting information. Graphics can be very effective in supporting text, illustrating mathematical concepts and increasing item access. If used improperly, however, graphics can add substantial confusion and distract test takers from the intent of the item. Graphics should be used judiciously.

- **Clarity** - Visuals should be clear and precise. Adding a visual may clarify the measurement intent of an item.
- **Mathematical Accuracy** - Visuals should utilize standard mathematical notation and formatting.
- **Simplicity** - Visuals should only contain necessary information. Remove unnecessary graphics. Avoid misleading graphics, such as charts with inconsistent scales.
- **Completeness** - Visuals should provide a representation of the important parts of the item. Visuals should mirror and parallel the wording and expectations of the problem.
- **If a visua**l within a given item is adding to the unintended difficulty of the item, it should be altered or removed.

Appropriate Use of Context

Contextual information includes problem scenarios, explanations, specific directions, and background text. Using contextual information can place mathematical concepts in more realistic conditions and provide background information that test takers may need. However, the contextual information should not interfere with the mathematics being assessed. It should not be a barrier to a student's ability to demonstrate his or her mathematical knowledge. Contextual information should be included only if the item is intended to assess mathematics in context.

- **Plain Language** - Use plain language as much as possible.
- **Increased Clarity** - Use manipulatives and/or graphics to increase item clarity.
- **Use Relevant Contexts** - Use context only if it is meaningful to the mathematics being assess.
- **Provide Appropriate Contexts** - Use contexts that are appropriate for the grade level being assessed.
- **Use Familiar Contexts** - Avoid contexts that may confuse or be unfamiliar to some students taking the assessment.

- **Provide Accurate Contexts** - Avoid contextual information that could interfere with the measurement of the intended skill.

Extraneous Information

Extraneous information includes all portions or aspects of an item that are unessential to the mathematics being assessed. This includes any inconsequential context. Extraneous information should be eliminated from all items in an accessible block.

- **Eliminate Extraneous Information**.
- **Provide Manipulatives Judiciously** - Only provide manipulatives when absolutely necessary (e.g., It may or may not be appropriate to test students' ability to visualize information using manipulatives).
- **Consider Item Context** - Provide students with units of measure only as necessary or appropriate for the context of an item. Including units of measure can be unnecessarily confusing.
- **Calculator Usage** - Do not ask students if they used a calculator for an item that obviously does not require its use.

Cues

Cues are components of item construction which give key information related to the problem. Cues can also give information related to incorrect answer choices. Cues can serve to clarify the intent of an item. Item writers should carefully consider how cues are used in each item.

- **Provide Descriptive Titles** - Identify the goal or topic of a problem with a title when appropriate. This is especially helpful for presenting word problems that require multiple pieces of information.
- **Provide Visual Cues** - **Bold**, *italicize*, <u>underline</u>, or CAPTIALIZE key words and phrases including:
  - Directions (e.g., <u>Solve</u>, COMPUTE, **Explain**) - Directions should <u>always</u> come at the **beginning** of a problem.
  - Operational words and phrases (e.g., **Add**, *Subtract*, <u>Find the product</u>)
- **Clarify Answer Requirements** - Cue students about the number and type of solution(s) they should provide (e.g., written description, graphical representation, etc.). This is especially important in open response items that could be solved using multiple approaches.
- **Avoid Deceptive Cues** - Do not mislead students to perform inappropriate operations.
- **Provide Definitions When Appropriate** - It may be appropriate to provide a brief definition, example, or illustration of a mathematical concept if doing so does not compromise the objective of the assessment item.

- **Use Cues to Clarify Item Intent** - Remember, the objective and intent of all testing items should be as clear as possible.

Computational Appropriateness

Each item on the NAEP is assigned a mathematical complexity rating (low, moderate, high). The task asked of the student should reflect an appropriate computational level. Generally, it is possible to reduce the computational complexity of an item while preserving its alignment with the NAEP framework.

- **Assess Computational Complexity** - Do not require students to perform calculations that are unnecessarily difficult. Calculations should not distract from the general idea being assessed in any given item.
- **Gauge Time Constraints** - Do not require students to perform calculations that are unnecessarily time consuming. Calculations should not distract from the "flow" of the testing experience. Remember that TIME is a precious resource during the testing experience.
- **Computational Progression** - Do not require students to perform counterintuitive operations.
- **Encourage Mathematical Accuracy** - Do not ask students to estimate or round when exact calculation is appropriate or easier.
- **Calculator Use** - Items should be constructed with calculator use/availability in mind. Computational complexity should be appropriate to the testing context. Remember, the availability of a calculator should not increase the complexity of a problem.

Grade Level Appropriateness

Item modifiers should identify the objective(s) being assessed by each NAEP item as well as the grade level at which it is meant to be assessed. Items in an accessible booklet should be constructed to assess at or below the grade level under consideration. For example, a fourth grade accessible block should not contain items that are constructed to assess a learning objective at an eighth grade level. In most cases, the NAEP framework provides leveled descriptions of each learning objective. Students being assessed using an accessible block should not be asked to perform a task at a level higher than is appropriate for their grade.

Cognitive Demand

Cognitive demand is a term used to refer to the overall difficulty of an item. Several components contribute the cognitive demand of any given item. For the purpose of creating an accessible block item writers should carefully consider factors which may unnecessarily increase the cognitive demand of an item.

- **Assessing Multiple Objectives** - Assessing multiple objectives in a single item generally increases the cognitive demand of an item. An accessible block should limit the number of items that assess multiple objectives.

- **Multiple Steps** - When possible, reduce the number of steps required to correctly answer an item while preserving the integrity of the objective being assessed.
- **Multiple Answers -** Limit the number of items that require multiple answer components.

Item Modification Procedures

1. Begin with a pre-developed (i.e., source) block of NAEP assessment items. There are several potential benefits to working with a source NAEP Block:
   - The block meets NAEP standards.
   - There may be information regarding item difficulty (e.g., % of students correctly answering the item, % of students selecting each alternative answer choice).
   - It may be possible to compare field test results with existing data.

2. Thoroughly review the document titled *Item Modification Guidelines*.
   - Each member of the item modification panel should have sufficient time to read and discuss the *Item Modification Guidelines*. The team should be presented with sample comparisons of original NAEP items with modified NAEP items and then be allowed to "practice" applying the recommendations on a few released NAEP items. This conversation should allow team members to become more comfortable and familiar with item modification guidelines and procedures.

3. Each member of the item modification panel should be given approximately 30 minutes to perform an initial individual review of each block. That is, each panel member should spend a short amount of time reading over each item, familiarizing themselves with the block. During this review each panel member should note:
   - Item and block clarity.
   - The diversity of NAEP objectives assessed by the block.
   - The difficulty of the items (% correct, % for each distracter).
   - Issues related to item quality (e.g., Are there errors? Do some items seem awkward or inappropriate for the grade level under consideration?).
   - Issues related to SD and ELL students (e.g., vocabulary and wording), particularly the use of calculators and manipulatives.
   - The balance of multiple choice and short response items.

4. The item modification panel should briefly discuss their thoughts from the initial item review. This conversation should be relatively brief (15-20 minutes). The following questions may be used to guide discussion:
   - Are there concerns about block or item clarity?
   - Is the block balanced? Is there a broad range of NAEP objectives assessed by the block, or are some learning objectives over or under represented by the block?
   - Are accessibility concerns effectively addressed?
   - Is the use of manipulatives/calculators necessary/appropriate?
   - Are all instructions clear?

5. Item-by-item review:

    - Modify each item as a group. It may be beneficial to have a large display of the item under consideration (i.e., use a projector).
    - Identify issues and concerns regarding the formatting, context, and accessibility of each item.
    - Carefully consider/modify the cognitive demand of each item. Each item must be addressed on a case by case basis, and considered in the context of the block as a whole. It is generally useful to refer to information regarding item difficulty (e.g., % of students correctly answering the item, % of students selecting each alternative answer choice) for this task.
    - Carefully review and apply the *Item Modification Guidelines*.
    - Record/Comment on recommended modifications to each item for future reference.
    - Record and classify the types of modifications that are recommended for each item. Use the document titled "Item Modification Record" to complete this process for each item.

    Note: It takes an average of 20-30 minutes to review each item. However, some items require less time to review (15 minutes) and some items required more time to review (50 minutes).

6. Compile all item modification recommendations. It is helpful to have each member of the panel submit their modified version of each item to the panel coordinator. Doing so often reveals misunderstandings or misinterpretations of group decisions regarding item modification. It is also helpful to have the group select one version of each modified item to serve as the representative sample of the panel's work for that item. It is convenient to use this representative sample of modified items as a reference for future editing and review procedures. It may be necessary to create an "editor ready" (i.e., clean copy) of some of the items.

7. Re-review all items in the block.

    - Note the degree of item modification on the Block Summary Sheet. Please refer to the document titled "Item Modification Rating Scale" for this task. This scale describes three levels of item modification, which may be useful for characterizing the overall degree of block modification.

APPENDIX E

ITEM RATING SCALE

Directions for Item Review Task

Thank you for agreeing to assist us with the item review process. If you have any questions about the item review process please feel free to contact Jeremiah Johnson at jeremiahmatthewjohnson@yahoo.com or (217) 714-6774.

Your packet of materials contains:

1. Seven blocks of modified 4th grade NAEP assessment items. These blocks have been modified systematically, according to a set of criteria that were developed for the purpose of this study.
2. Seven generic block coversheets (one for each block).
3. A document titled "Guidelines for Item Modification"
4. A document titled "Aligning the 'Accessible Block' Assessment with the NAEP Framework"
5. A document titled "Item Rating Scale"

We would like you to complete three tasks:

1. **For each item, rate the adequacy of the mathematical content being assessed**. This process is similar to the task you completed in Boston in February. We would like you to rate each item using the "Item Rating Scale" (attached). **Please write your rating (1-3) and brief comments related to item content on the block cover sheet.** If you have ideas or suggestions for editing or improving any of the proposed items in a block, please feel free to write your ideas directly on the item. A team will thoroughly review all of your comments and suggestions.

   If you wish, you may also comment on the construction of the block as a whole (e.g., balance of mathematical concepts assessed within a block, general item difficulty level, ideas for improving block format, etc. These comments can also be noted on the block cover sheet.

2. **For each block, comment on general item/block alignment with the modification specifications provided in the document titled "Guidelines for Item Modification"**. **Your comments may be written on the block cover sheet.**

3. **For each block, comment on item/block alignment with the NAEP framework.** The 2005 NAEP mathematics framework is available online at: http://www.nagb.org/pubs/m_framework_05/761607-Math%20Framework.pdf

It may be helpful to refer to the document titled "Aligning the 'Accessible Block' Assessment with the Framework" for this task. More specifically, we would like you to comment on the match between the emphasis of the assessment and the emphasis of topics,

objectives, and contextual requirements in the framework: The assessment should represent the balance of content and item formats specified in the framework and give appropriate emphasis to the conditions in which students are expected to demonstrate their mathematics achievement.

Item Rating Scale

Each NAEP item should assess mathematical content. In addition, items should assess the student's ability to reason with the content. The assessment should give all students tested a reasonable opportunity to demonstrate their knowledge and skills in the topics and objectives covered by the framework (with a special emphasis here being placed on providing students at the lower end of the achievement spectrum an opportunity to show what they know and are able to do).

PLEASE RATE THE MATHEMATICAL ADEQUACY OF EACH ITEM USING THE FOLLOWING SCALE.

1. **Adequate**
   The problem is posed clearly. Any student who learned the mathematics of the task should be able to understand what is being asked. There are no unreasonable hidden assumptions. The context, language, and/or graphics used to pose the problem do not create unnecessary challenges that are unrelated to the mathematics. The problem, along with its response set or scoring rubric, does not contain mathematical errors.

2. **Marginal**
   The item is somewhat problematic. It may work as intended for may students, but defects in the item may unnecessarily lead to error or frustration for some students. In some cases, a simple edit may be sufficient to render the item adequate.

3. **Seriously Flawed**
   Item fails substantially on one or more of the following criteria: a) it is undermined by hidden assumptions that are unfair to the student; b) the context is confusing and misleading in ways that are not related to what is being measured; c) the language and graphs present unnecessary obstacles to understanding what is being posed; or d) their are mathematical errors in the problem or in its response set or scoring.

APPENDIX F

COGNITIVE LAB GUIDE

Cognitive Lab Guide

Purpose:

The purpose of this cognitive lab is to gain insight into how students interpret and respond to items. During the cognitive labs, purposive samples of approximately 5 students will take both an "accessible block" and a "standard block" in a counterbalanced design using a 1:1 administration with a trained observer. The observer will prompt the student to "think aloud" as they complete the item blocks and debrief the student as to strategies used. Student work will also be analyzed to identify strategies and evaluate performance. Comparisons will be made between strategies, time to completion and performance across "accessible" and standard blocks.

Procedure:

1. Review the Student Assent Form and have all students sign this form prior to beginning the cognitive lab.
2. Describe the purpose of the cognitive lab to the student.
3. Begin the administration of the testing blocks. Follow the standard testing procedures specified by NAEP as closely as possible (e.g., read all directions aloud, only use testing aids such as calculators when permitted by the test, etc.)
4. Ask students to "think aloud" as they complete items contained in the assessment blocks.
5. Ask probing question when appropriate, but try not to disrupt the flow of the testing process more than necessary.
6. Conclude each cognitive lab by thanking the student for his or her participation.

Examples of questions that may be asked during the cognitive lab:

- Are the directions clear? If not, why not?
- Is the question clear? If not, why not?
- Is any part of this problem confusing? What part?
- Are any of the words in this problem hard to read or understand?
- Is the image/graph/picture used in this problem clear?
- What are you thinking?
- Is this problem difficult? What is the most difficult part of this problem?
- Have you eliminated any of the answer choices? Why?
- Why did you skip that problem?